University of Groningen

Introductory Essay

Evolutionary Changes in Data Analysis

Author:
Mostafa Hadadian Nejad Yousefi

First Promoter:
Alexander Lazovik

Second Promoter:
Viktoriya Degeler

April 7, 2021

Contents

List of Tables

List of Figures

1 Introduction

Data is the new oil [1]. We heard this a lot since 2006. Although this comparison is not accurate, we can still use it to define data analysis. They are similar in that we should first extract them and then process them to gain valuable products. Data analysis is the process of converting raw data to worthy outcomes. It is a multi-disciplinary field requiring skills ranging from mathematics, statistics, and computer science to business understanding in different areas.

Data and oil are different so that, unlike oil, data is not depleting in the process, and also it comes in different forms. Data could be any information and fact such as water quality, traffic control picture, customer service recordings storing in different types such as numbers, video, and sound, respectively. The most astonishing privilege of data over oil is its usefulness in almost every businesses.

2 Motivational Example

To better understand the ECiDA project's motivations, we present an example from various real-world data analysis applications. Further will use this example through this document to explain each section.

The ever increasing development in data science enables industries to use data-centric projects ranging from healthcare [2] to manufacturing [3]. Predictive maintenance is one of these projects for preventing failure in productions, especially in Industry 4.0. Different applications are defined in this field, such as predicting the remaining useful life of a machine using data analysis techniques can avoid damage to the industry by replacing the device at the right time also to minimize the operating expenses [4].

Figure 1 shows an overview of a sample predictive maintenance pipeline derived from literature in related field [5, 6, 7]. This pipeline collects streaming data from sensors and analyzes data to extract useful insights for maintenance use cases. Besides, a batch processing pipeline receives all data points and evaluates the streaming pipeline's performance instead of analyzing only the events. Using feedbacks from batch processing, we should update the stream processing pipeline to improve the accuracy, e.g. by capturing new events. It is critical to have the streaming pipeline up and running 24x7 and to deliver the updates as soon as possible to prevent irrecoverable damages. For example, a malfunction in a boiler or toxic water could put lives in danger, while a timely update in the stream processing pipeline could prevent the catastrophe.

The more data and the more variant data types improve the predictive model's potential to reach better performance. Various sources are used for predictive maintenance, visual inspections such as thermal imaging, physical conditions such as noise generated by the device, internal device states. Some may want to start with visual inspection first and then add other sensors to make the model more accurate over time. In fig. 1, the input sources are represented as hexagons. Each sensor is placed on several devices. For example, the thermometers are installed on boilers, motors, and pumps.

Event detection components are on the next layer. They collect data from sensors and find pre-defined patterns in data to capture unique events. Extracting useful information in the first layer reduces network usage and response time. For instance, a thermometer is installed on a motor, and the heat detector captures events when the temperature is above a threshold. Another instance is an event detector looking for an event when the device is rattling and loud to anticipate motor malfunction. Since there are different devices, we may need to have distinct event detectors for each device types, e.g. the threshold differentiates between a motor and pump.

Next, the aggregator collects data points from different (in type and device) sources and order them chronologically, then joins them by device and timestamp such that each record represents every event that occurred at a specific time on a particular machine. In this component, input data points are dependent as they formed an output record together.

The desired applications are placed in the next layer, where all components receive data from the aggregator. Each component works independently and analyzes data to produce the desired outcome. For instance, the failure prediction can predict an upcoming malfunction happening shortly if the motor temperature unexpectedly goes high while it is rattling and generating intense noise.

Whenever a result is ready, the applications will send it to the end-point that is responsible for acting on the insights. If there is a critical issue in the company, instantly act upon that, like shutting down the motor. Otherwise, alarm the operator. Although this component receives three distinct inputs, the inputs are independent, and actions are defined per input.

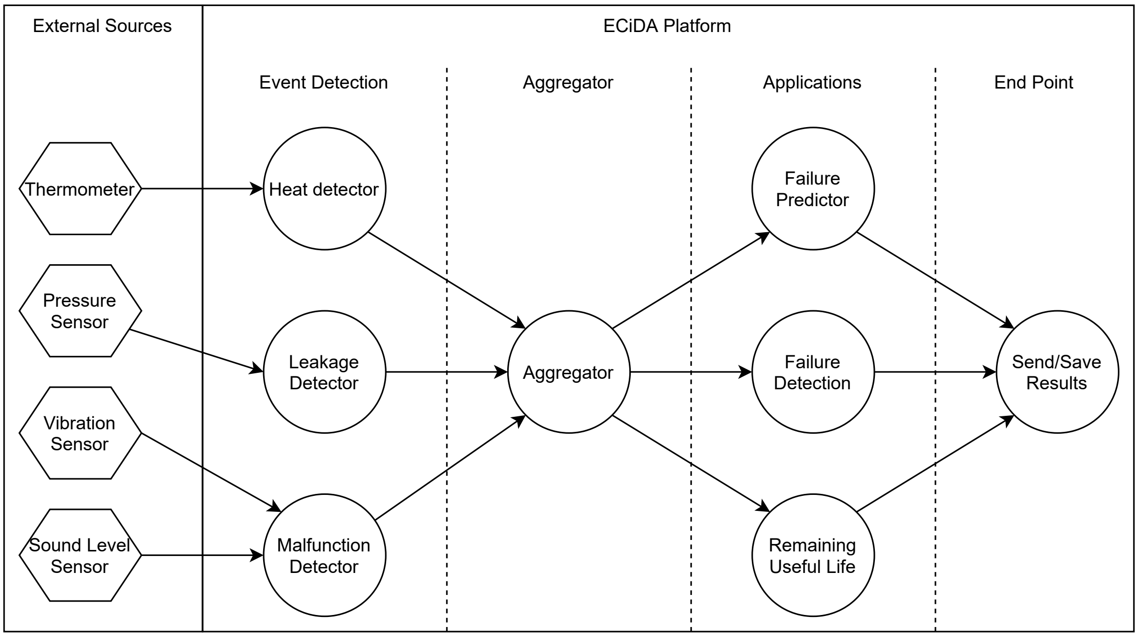

Figure 1: **A predictive maintenance pipeline:** Circles are controllable components inside the ECiDA platform, while the hexagons are external sensors.

As mentioned above, a batch processing application runs periodically to analyze the whole data (not just events) and also evaluate the streaming pipeline performance. We can update the real-time pipeline according to the insights from the entire data point set and the actual events that happened in the real world, which can assess the predictions. A resulting update can happen in any component.

We may want to update the currently running model in the application layer to a newly trained model. We may also want to capture new events by adding more event detectors leading to updating the aggregator. We may even find that an event is of no use in the prediction and only occupied resources, and it is better to be removed from the pipeline. The change might be subtle such as changing the threshold, or it could be as immense as changing the entire pipeline, including components and connections.

References

[1] C. Humby, "Data is the new oil," *Proc. ANA Sr. Marketer's Summit. Evanston, IL, USA*, 2006.

[2] Y. A. Qadri, A. Nauman, Y. B. Zikria, A. V. Vasilakos, and S. W. Kim, "The future of healthcare internet of things: a survey of emerging technologies," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1121–1167, 2020.

[3] A. Ismail, H.-L. Truong, and W. Kastner, "Manufacturing process data analysis pipelines: a requirements analysis and survey," *Journal of Big Data*, vol. 6, no. 1, pp. 1–26, 2019.

[4] R. K. Mobley, *An introduction to predictive maintenance.* Elsevier, 2002.

[5] R. Sahal, J. G. Breslin, and M. I. Ali, "Big data and stream processing platforms for industry 4.0 requirements mapping for a predictive maintenance use case," *Journal of manufacturing systems*, vol. 54, pp. 138–151, 2020.

[6] Y. Yamato, H. Kumazaki, and Y. Fukumoto, "Proposal of lambda architecture adoption for real time predictive maintenance," in *2016 fourth international symposium on computing and networking (CANDAR)*, pp. 713–715, IEEE, 2016.

[7] M. Short and J. Twiddle, "An industrial digitalization platform for condition monitoring and predictive maintenance of pumping equipment," *Sensors*, vol. 19, no. 17, p. 3781, 2019.