Chapter 12
Process Mining in the Large

Process mining provides the technology to leverage the ever-increasing amounts of event data in modern organizations and societies. Despite the growing capabilities of modern computing infrastructures, event logs may be too large or too complex to be handled using conventional approaches. This chapter focuses on handling "Big Event Data" and relates process mining to Big Data technologies. Moreover, it is shown that process mining problems can be decomposed in two ways, *case-based decomposition* and *activity-based decomposition*. Many of the analysis techniques described can be made scalable using such decompositions. Also other performance-related topics such as streaming process mining and process cubes are discussed. The chapter shows that the lion's share of process mining techniques can be "applied in the large" by using the right infrastructure and approach.

12.1 Big Event Data

Some of the process mining tools described in Chap. 11 can discover process models for logs with billions of events. However, performance highly depends on the characteristics of the event log (e.g., number of distinct activities and redundancy) and the questions asked (e.g., conformance checking is often more time consuming than discovery). Moreover, for some applications event logs may be even larger or results need to be provided instantly.

In Chap. 1, we listed the "four V's of Big Data": Volume, Velocity, Variety, and Veracity (Fig. 1.4). The term "Internet of Events", introduced in Sect. 1.1, refers to the growing availability of event data. These data are omnipresent and an enabler for process mining. This chapter will focus on event data and the first two V's. The first 'V' (Volume) refers to the size of some data sets, in our case event logs. We will discuss various *decomposition* and *distribution* strategies to turn large process mining problems into multiple smaller ones. The second 'V' (Velocity) refers to the speed of the incoming events that need to be processed. It may be impossible or undesirable to store all data. Therefore, we will also introduce the topic of *streaming* process mining.

W. van der Aalst, *Process Mining*, DOI 10.1007/978-3-662-49851-4_12

Big Data is not limited to process-related data. However, Big Data infrastructures enable us to collect, store, and process huge event logs (see Sect. 2.5.9). Process mining tools can exploit such infrastructures. Therefore, we describe current trends in hardware and software (Sect. 12.1.2), before describing the characteristic features of event logs (Sect. 12.1.3). However, first we briefly discuss the possibilities and risks when going from sampled "small data" to "all data".

12.1.1 N = All

In the past, conclusions were often based on human judgment or analysis of sample data. Either data were not available, unreliable, or it was impossible to process all data. In many businesses, we now witness a change from collecting *some data* to collecting *all data* [100]: "$N = \text{All}$" where N refers to the sample size. As described in Sect. 1.1, the digital universe and physical universe are becoming more aligned. Money has become a predominantly digital entity. Queries on the availability of products are answered based on data in some database rather than a visit to the warehouse. The direct coupling between data and reality combined with our improved abilities to store and process data forms the playground of data science as described in Chap. 1.

Sampling was needed in the "analog era" characterized by information *scarcity*. Due to sampling error and sampling bias, it may be risky to extrapolate conclusions from sample data. Moreover, the granularity of analysis using sampled data is often too coarse making it impossible to draw conclusions for smaller subcategories and submarkets. Hence, the concept of sampling makes no sense if *all* data are available and we have the computing power to analyze all events. Consider, for example, conformance checking. Why just check the conformance of a few cases if we can check all cases and detect all deviations? Clearly, "$N = \text{All}$" requires a new way of thinking. For example, auditors and accountants may be afraid of uncovering all deviations. Also the work of marketers and social scientists is changing: large-scale data analysis is replacing sampling and questionnaires.

Having all data ($N = \text{All}$) may also create problems:

- Hardware, software and analysis techniques need to be able to cope with the associated volumes.
- Overfitting the data may lead to "bogus conclusions" (cf. Bonferroni's principle).

This chapter will focus on the first problem. However, to illustrate the second problem we consider the following example inspired by a similar example in [114].

Suppose some Dutch government agency is searching for terrorists by examining hotel visits of all of its 18 million citizens (18×10^6). The hypothesis is that terrorists meet multiple times at some hotel to plan an attack. Hence, the agency looks for suspicious "events" $\{p_1, p_2\} \dagger \{d_1, d_2\}$ where persons p_1 and p_2 meet on days d_1 and d_2 in some hotel. How many of such suspicious events will the agency find if the behavior of people is completely random? To estimate this number, we make some