

Increasing Scalability of Process Mining using Event Dataframes: How Data Structure Matters

Alessandro Berti¹ [0000–0003–1830–4013]

Process and Data Science department, Lehrstuhl für Informatik 9 52074 Aachen,
RWTH Aachen University, Germany

Abstract. Process Mining is a branch of Data Science that aims to extract process-related information from event data contained in information systems, that is steadily increasing in amount. Many algorithms, and a general-purpose open source framework (ProM 6), have been developed in the last years for process discovery, conformance checking, machine learning on event data. However, in very few cases scalability has been a target, prioritizing the quality of the output over the execution speed and the optimization of resources. This is making progressively more difficult to apply process mining with mainstream workstations on real-life event data with any open source process mining framework. Hence, exploring more scalable storage techniques, in-memory data structures, more performant algorithms is a strictly incumbent need. In this paper, we propose the usage of mainstream columnar storages and dataframes to increase the scalability of process mining. These can replace the classic event log structures in most tasks, but require completely different implementations with regards to mainstream process mining algorithms. Dataframes will be defined, some algorithms on such structures will be presented and their complexity will be calculated.

Keywords: Process Mining · Event Logs · Big Data · In-Memory Processing.

1 Introduction

Information systems support the execution of nowadays business processes, recording information about *events* related to the executions of a business process. Relational (MySQL, Oracle, SQL Server, ...) and non-relational databases (MongoDB, Elasticsearch, ...) are used to store and query the event data. Process Mining supports the extraction of process-related knowledge from information systems, targeting both systems supported by a workflow (some possible analyses are *bottleneck analysis*, *remaining time prediction*, *logical-temporal checking*) and systems where process executions are more loosely described (with *process discovery*, that is the automatic discovery of a process schema from the event data, and *conformance checking* techniques). Several process mining algorithms have been developed for each goal. Generally, a process mining analysis starts from the extraction of an *event log* from the information system. The event log

Table 1. Characterization of the event logs used in the assessment of the approaches. The loading time and the occupation of RAM have been measured against the XESLite log importer in the ProM6 process mining framework. The measurement has been done on a Lenovo ThinkPad T470s notebook (I7-7550U, 16 GB RAM, DDR).

Log	Events	Cases	Variants	Classes	Loading time(s)	RAM(MB)
roadtraffic	561470	15370	231	11	9.60	412
bpic2017_o	193849	42995	16	8	6.70	421
bpic2017_a	1202267	31509	15930	26	49.9	1203
bpic2018	2514266	43809	28457	41	39.4	3219
bpic2019	1595923	251734	11973	42	29.8	1435

is then imported into a process mining tool (for example, the ProM6 Process Mining framework), and several analyses to the same log can be applied. As an example, a popular process discovery technique is the inductive miner [12], that extract a process model (process tree or accepting Petri net) from the event log. Alignments are a popular conformance checking technique [4], that enables compliance and performance analyses.

A problem that surged in recent years is that the amount of information contained in modern information systems is steadily increasing. This pose challenges that were addressed only from a small part of the process mining research. Taking into account a mainstream analytic workstation in 2019 (I7-8550U, 16GB DDR4, 3000 MB/s SSD sequential read), and a public available log as the BPI Challenge 2018 log (2.5M events, 44K different process executions), importing the log through the efficient XESLite importer takes 50 seconds and the resulting in-memory log structure occupies 3219 MB of RAM. More details about the features of the considered logs are contained in section 1. On the other hand, in recent years some efficient storage techniques have been proposed, that are called *columnar* because each column is saved separately inside the storage, including the popular Apache Parquet and Apache ORC formats. Corresponding to the storage techniques, an interesting class of data structures, the *dataframes*, has been implemented to support fast filtering and aggregations of the underlying data. With regards to open source process mining, at the best of our knowledge no previous usage of columnar storages in process mining has been described, while a previous scientific contribution [9] has used a dataframe structure in order to increase the speed of computation in process mining, but without providing an abstraction of the data structure, or some considerations on why dataframes are better suited to handle big amount of data.

In this paper, an abstraction of the notion of *dataframe* is provided (Pandas and Spark provide implementations of such structure), explaining a basic implementation of dataframe, with the final goal to obtain a greater scalability in mainstream process mining operations. Dataframes can directly ingest event data contained in columnar storages (like Apache Parquet and Apache ORC), and so they can exploit their advantages (attribute selection at load time and efficient compression techniques). Some complexity estimation for mainstream process mining operations, like filtering and the Directly-Follows Graph computation, on the described dataframe implementation will be provided in order

to show how dataframes are suited for fast log exploration and filtering. At the best of the authors knowledge, this is the first scientific paper combining in an holistic manner columnar storages and dataframes to increase the scalability of process mining techniques. Moreover, this is (at the best of the authors knowledge) the first abstraction of the dataframe concept, that has already very popular implementations in Pandas and Apache Spark. The assessment (while avoiding, for fairness, direct comparison with the implementation of the classical event log structure, done in different frameworks/languages) will show that popular dataframe implementations can handle big amount of data with a good execution time and a reasonable memory occupation.

The paper is organized as follows. In section 2, the related work on storage techniques and in-memory/distributed processing and benchmarking is presented. In section 3, the concepts necessary to understand the paper will be introduced. In section 4, the complexity of the operations on classic event logs will be presented. In section 5, the dataframe concept will be introduced along with some transformation functions, a conversion technique from an event log structure to a dataframe structure, a simple implementation and a calculation of the complexity of operations on dataframes. In section 6, a dataframe implementation in the Python language (Python) will be assessed on real-life and simulated event logs. Eventually, section 7 concludes the paper.

2 Related Work

The XES standard: several formats have been proposed during the years for the standard storage of event logs in process mining. The IEEE standard is XES [19], for which different implementations exist in the ProM6 process mining framework. Among noticeable implementations, we can cite XESLite [16], that provides a memory-efficient handling of event logs, while [10] supports the XES standard on relational databases, albeit with a performance deficit, and DB-XES [17] that use relational databases to support some intermediate calculations. OpyenXES [18] took XES in an open source Python implementation, and the PM4Py Python process mining library [8] followed obtaining a full certification.

Benchmarking of process mining algorithms: the work [5] contains an extensive benchmarking of some process mining algorithms. The evaluation highlight some problems of available techniques, including the lack of scalability and a strong divergence in their performance with respect to the different quality metrics. In [13], a variant of the popular inductive miner process discovery algorithm (based only on the directly-follows graph) is explained, along with extensive evaluation.

Process Mining over Big Data: an approach of process mining over big data triples stores is contained in [6]. A ProM 6 plug-in, enabling process mining on the Apache Hadoop big data platform, has been described in [11]. In [1], some motivations on connecting process mining with the big data world are explained. A distributed compliance monitoring of business processes over Mapreduce architectures is described in [15]. An approach of process mining over the Apache

Spark framework is described in [9].

Row- and Columnar- based approaches in Data Mining: in [3], a comprehensive evaluation of column-stores versus row-stores is provided. The implementation of a modern column-oriented database system is described in [2].

3 Background

3.1 Event Logs: Definition and Considerations

In this section, the notion of event log, i.e., a collection of events recorded in the execution of a business process that is the starting point of many process mining analysis, is formally introduced. For the following definition, let \mathcal{U}_E be the universe of events, \mathcal{U}_C be the universe of case identifiers, \mathcal{U}_A be the universe of activities (names referring to a particular step of a process), \mathcal{U}_{attr} be the universe of attribute names (all the names of the attributes that can be related to an event), \mathcal{U}_{val} be the universe of attribute values (all the possible values for attributes).

Definition 1 (Classical Event Log) *A log is a tuple $L = (C_I, E, A, case_ev, act, attr, \leq)$ where $C_I \subseteq \mathcal{U}_C$ is a set of case identifiers, $E \subseteq \mathcal{U}_E$ is a set of events, $A \subseteq \mathcal{U}_A$ is the set of activities, $case_ev \in C_I \rightarrow \mathcal{P}(E) \setminus \{\emptyset\}$ maps case identifiers onto set of events (belonging to the case), $act \in E \rightarrow \mathcal{U}_A$ maps events onto activities, $attr \in E \rightarrow (\mathcal{U}_{attr} \dashv \mathcal{U}_{val})$ maps events onto a partial function assigning values to some attributes, and $\leq \subseteq E \times E$ defines a total order on events.*

This classical event log notion matches the XES storage format [19], that is the common source of information for process mining tools like Disco, ProcessGold, Celonis, QPR, Minit, ... An example attribute of an event e is the timestamp $attr(e)(time)$ which refer to the time the event happened. While, in general, an event belongs to a single case, in Def. 1 the function $case_ev$ might be such that cases share events.

An important consideration is that for an event $e \in E$, applying $attr$, the resulting mapping $\mathcal{U}_{attr} \dashv \mathcal{U}_{val}$ is a partial function. This means that not all the attributes are valued for all the events, e.g., there might be some attributes that are specific for a single event.

3.2 Directly-Follows Graph

The Directly-Follows Graph (DFG) is a basic representation of the activities of the process, that are connected with a directed if they follow each other in at least an instance of the event log. The DFG is usually annotated by the number of occurrences for each edge (corresponding to the number of occurrences of the log where an activity has been followed by another activity in the log).

The concept of a Directly-Follows Graph is formally introduced in the following definition:

Partial loading (some attributes)

S1	case:concept:name	1	2	1	2	1	1	2	2
S2	concept:name	A	A	B	B	C	D	C	D
S3	time:timestamp	1970	1971	1972	1973	1974	1975	1976	1977

Complete loading (all attributes)

S1	case:concept:name	1	2	1	2	1	1	2	2
S2	concept:name	A	A	B	B	C	D	C	D
S3	time:timestamp	1970	1971	1972	1973	1974	1975	1976	1977

Fig. 1. Example selective loading of attributes. For most process discovery and conformance checking algorithms, loading only the case ID (*case:concept:name*) and the activity (*concept:name*) is enough. So, why loading all the attributes if they are not needed? For log exploration purposes, instead, it may be useful to load all the attributes (in order to filter interesting behavior).

Definition 2 (DFG) A *Directly-Follows Graph* is a weighted directed graph: $DFG = (N_{DFG}, E_{DFG}, c_{DFG})$ where N_{DFG} (the nodes) are the activities, and $E_{DFG} \subseteq N_{DFG} \times N_{DFG}$ is the set of all the edges between activities that happened in direct succession, and $c_{DFG} : E_{DFG} \rightarrow \mathbb{R}^+$ is the count function that aims to represent how many times two different activities happened in direct succession.

Let $start_{DFG} \subseteq N_{DFG}$ be the subset of start activities in the log, and $end_{DFG} \subseteq N_{DFG}$ be the subset of end activities in the log. Then, the DFG structure can be converted into a Petri net useful for conformance checking purposes using the method described in [14]. Moreover, a variant of the Inductive Miner (IMDF) is described in [13].

3.3 Row- and Columnar-based Storages

This section will introduce row- and columnar-based storages. While the remainder of the paper will be mainly focused on columnar storages and data structures, it is interesting to see how modern row-based storages cope against the columnar storages when saving process mining event logs.

Row-based storage systems are optimized to read logs “by rows”, “by events”: while parsing, the row is considered in its entirety, and all the values for all the attributes of the parsed event are read at first glance. A well-known row-based storage format is the Avro format¹. Each Avro file contains an header, where the parsing schema is defined, and one or more data blocks (containing one or more rows). For process mining purposes, the event log can be transformed into a list of events, and then each event serialized into a row.

Column-based storage systems are optimized to read logs by “by columns”, making possible to choose which attributes are needed before starting to load

¹ <https://avro.apache.org/>

Table 2. Size on disk of logs stored in XES, CSV, Avro, Parquet (with Snappy compression), ORC, XES (applying Gzip compression to the file), Parquet (applying Gzip compression to the columns). The Gzip compression leads to a performance hit since the file and the columns need to be uncompressed before loading. It should be noted that, while in the CSV/ORC/Parquet formats only the event data is contained in the file, XES contains some additional metadata and information.

Log	S.XES	S.CSV	S.Avro	S.Parq.(S)	S.ORC	S.XES(G)	S.Parq.(G)
roadtraffic	181MB	47MB	34MB	6MB	3.9MB	3.3MB	3.6MB
bpic2017_o	107MB	36MB	34MB	7MB	5.5MB	5.5MB	4.7MB
bpic2017_a	565MB	205MB	184MB	29MB	22.0MB	29.1MB	17.1MB
bpic2018	1901MB	1505MB	1955MB	167MB	146MB	155MB	96MB
bpic2019	711MB	434MB	462MB	14MB	10.1MB	16.5MB	8.6MB

the file (see Figure 1). In this way, only the values of some attributes are parsed, making the load operation faster. Since each column of the file contains data of the same format (integer, string, ...), more effective compression techniques can be deployed. Among the compression algorithms supported by Parquet, there is the *Snappy* compression² and the *Gzip* compression³. Snappy is more fast in compressing/decompressing, while Gzip obtains better compression ratios at the expense of performance. Apache Parquet⁴ is supported by a large number of big data frameworks (Apache Hive, Apache Drill, Apache Impala, Apache Pig, Apache Spark, Cascading, ...). Apache ORC⁵ is an alternative in the columns. In Figure 3.3, the size on disk of the logs stored in the different formats is reported. The Apache Parquet format with the Gzip compression achieves the best results in terms of log size on disk, while Parquet with the Snappy compression remains very efficient while avoiding the compression/decompression performance deficit of Gzip.

4 Approach: Complexity of operations on Attribute Maps

The aim of this section is to present the complexity of retrieval and of the Directly-Follows Graph calculation on top of classical event logs. This will permit to compare the complexity of such operations against an implementation of the dataframe, that is introduced in section 5.

In section 2, the current standard-de-facto implementations of event logs, based on the XES standard, have been presented. The *attr* function is usually implemented to return $\mathcal{U}_{attr} \not\rightarrow \mathcal{U}_{val}$ as an associative array data structure (for example, the plain XES uses the *HashMap* and XESLite uses the *THashMap* (Trove map implementation)). Hash maps are efficient in insertion and retrieval, with an average complexity of $O(1)$, but a worst case of $O(M)$ (where M is the

² [https://en.wikipedia.org/wiki/Snappy_\(compression\)](https://en.wikipedia.org/wiki/Snappy_(compression))

³ <https://en.wikipedia.org/wiki/Gzip>

⁴ <https://parquet.apache.org/>

⁵ <https://orc.apache.org/>

case id	activity	timestamp	resource	cost
1	A	1970-01-01	Mike	10
2	A	1971-01-01	Maria	20
1	B	1972-01-01	Tom	10
2	B	1973-01-01	Maria	30
3	A	1974-01-01	Luke	10
2	C	1975-01-01	Tom	15
3	B	1976-01-01	John	20

Fig. 2. Representation of a dataframe structure. In this table, the columns are corresponding to the names of the attributes.

number of attributes of the specific event). This means that the complexity of the retrieval of the entire set of values of a single event attribute, on all the events of the log, is in the worst case $O(N * M)$ (where N is the number of events in the event log and M is the maximum number of different attributes that any event has). The complexity in the average case is $O(N)$. This holds also for common problems in process mining like filtering the log on specific attribute values, or extracting the Directly-Follows Graph (see subsection 3.2).

The calculation of the Directly-Follows Graph could be performed iterating over the cases of the log. This has an average complexity of $O(N)$ since a single pass is enough and the storage of the edges/counts in a map on the log has an average linear complexity, but in the worst case (where all the events belongs to a single case and the worst case of a map is encountered) has complexity of $O(N^2)$.

5 Approach: Dataframes

This section introduces the techniques that are necessary to do process mining on top of a different structure, i.e. the *dataframe*. Dataframes can ingest without hassle data contained in popular columnar formats like Apache Parquet or Apache ORC, that were introduced in section 3.

The section is organized as follows: in subsection 5.1, the abstraction of the dataframe structure is presented. In subsection 5.2, a basic implementation of a dataframe is presented, along with some conversion techniques from an event log structure to a dataframe abstraction. In subsection 5.3, some basic operations (that are useful for process mining purposes) on a dataframe structure are introduced. In subsection 5.4, some algorithms to calculate the Directly-Follows Graph on dataframes are presented.

5.1 Definition of Dataframes

To introduce the concept of dataframe, one could think to a set of attribute names (columns) associated to the values (of the attribute) for each row (that is an event). An example representation is contained in Fig. 2.

Definition 3 (Dataframes) Let $I \subset \mathbb{N}$ be a set of indexes, $N \subset \mathcal{U}_{attr}$ be a set of attribute names, $T \subset \mathcal{U}_{types}$ be a set of attribute types, $V \subset \mathcal{U}_{val}$ be a set of attribute values for which a partial order \leq_V exists. A dataframe is a tuple $D = (I, N, T, V, \chi_i^{val}, \chi^{type})$ where:

- $\chi_i^{val} : I \times N \rightarrow V \cup \{\epsilon\}$ assigns to an index $i \in I$ and an attribute name $n \in N$ a value in $V \cup \{\epsilon\}$.
- $\chi^{type} : N \rightarrow T$ assigns to an attribute name $n \in N$ a type in T .

An assumption that is being made is that attributes n_{case} (the case) and n_{act} (the activity) exist, such that for all $i \in I$, $\chi_i^{val}(n_{case}) \neq \epsilon$ and $\chi_i^{val}(n_{act}) \neq \epsilon$.

In the following, we consider \mathcal{U}_D to be the *universe* of dataframes, and $\chi^{val} : N \rightarrow I \times (P \cup \{\epsilon\})$ is a function that associates to an attribute name the set of all indexed values, i.e., for $n \in N$: $\chi^{val}(n) = \cup_{i \in I} \{(i, \chi_i^{val}(n))\}$. Hence, if we take the second element of the couples, we get the set of all attribute values for n , that we indicate with $\pi_2(\chi^{val}(n))$. Moreover, given a dataframe $D \in \mathcal{U}_D$, the function $\pi_I : \mathcal{U}_D \rightarrow \mathcal{P}(\mathbb{N})$, $\pi_I((I, N, T, V, \chi_i^{val}, \chi^{type})) = I$ is defined.

5.2 Simple Implementation of a Dataframe, Conversion from Event Logs to Dataframes

The data structure defined in subsection 5.1 is actually an abstraction of the one that is implemented in Pandas (Python) and Apache Spark. In this section, we want to describe a simple implementation of the dataframe concept, assuming that the name of the attributes is represented as a string. The proposed implementation is the following: **Map**<**String**, **ArrayList**<**Object**>> where: **Map** is any associative array implementation⁶ and **ArrayList** is any dynamic array implemented⁷. In the previous data structure, all the values for a particular attribute can be retrieved from the map as an array list of objects. In this data structure, the value in position i is the value of the attribute for the event at position i in the dataframe.

Converting from an event log $L = (C_I, E, A, case_ev, act, attr, \leq)$ to a dataframe $D = (I, N, T, V, \chi_i^{val}, \chi^{type})$ requires the assumption that an attribute $n_{case} \in \mathcal{U}_{attr}$ exists in the domain of $attr(e)$ for each event $e \in E$. Since \leq (see 1) is a total order on E , E can be viewed as a sequence of events, $E = \{e_1, e_2, \dots, e_n\}$, such that $e_i \leq e_j$ if and only if $i \leq j$. The conversion is such that $N = \mathcal{U}_{attr}$ is the set of attributes of the dataframe, and $T = \{o\}$ can be defined to contain a single (object) type. Then, $V = \cup_{e \in E} \text{Im } attr(e)$, $\chi^{type}(n) = t \forall n \in N$, and

$$\chi_i^{val}(n) = \begin{cases} attr(e_i)(n) & \text{if } n \in \text{Dom } attr(e) \\ \epsilon & \text{if } n \notin \text{Dom } attr(e) \end{cases}$$

⁶ https://en.wikipedia.org/wiki/Associative_array

⁷ https://en.wikipedia.org/wiki/Dynamic_array

5.3 Transformation functions

We introduce the following functions in order to be able to perform process mining operations on top of a dataframe (as the calculation of Directly-Follows Graphs). These functions transform a/some dataframe(s) into another dataframe.

- *Projection on a given expression*: given a dataframe $D = (I, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})$, a set of attribute names $S \subseteq N$ and a function $f : I \times \mathcal{P}(N) \rightarrow \{0, 1\}$ (selective function based on the attribute values), a projection function is defined as $\text{proj}((I, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}}), S, f) = (I', N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})$ where $I' = \{i \in I, f(i, S) = 1\}$ and $\chi_i^{\text{val}}(n) = \chi_i^{\text{val}}(n)$ for all $i \in I'$ and $n \in N$.
- *Grouping function*: the grouping function is defined given a dataframe $D = (I, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}}) \in \mathcal{U}_D$ and an attribute $n_0 \in N$ as

$$\text{group}(D, n_0) = \bigcup_{v \in \pi_2(\chi^{\text{val}}(n_0))} \text{proj}(D, n_0, \{v\})$$

by construction, $\text{proj}(D, n_0, \{v\}) \cap \text{proj}(D, n_0, \{v'\}) = \emptyset$ if $v \neq v'$, so the sets are pairwise disjoint.

- *Shift function*: $\text{shift}((I, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})) = (I', N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})$ where $I' = \{i - 1 \mid i \in I\}$, $\chi_{i-1}^{\text{val}}(n) = \chi_i^{\text{val}}(n) \quad \forall i \in I, n \in N$.
- *Concatenation function*: for $D_1 = (I_1, N_1, T_1, V_1, \chi_i^{\text{val}}, \chi^{\text{type}}) \in \mathcal{U}_D$, $D_2 = (I_2, N_2, T_2, V_2, \chi_i^{\text{val}}, \chi^{\text{type}}) \in \mathcal{U}_D$, and $s \in \Sigma$ such that $\{n_2 \oplus s \mid n_2 \in N_2\} \cap N_1 = \emptyset$, the concatenation is defined as $\text{concat}(D_1, D_2, s) = (I, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})$ where $I = I_1 \cup I_2$, $N = N_1 \cup \{n_2 \oplus s \mid n_2 \in N_2\}$, $T = T_1 \cup T_2$, $V = V_1 \cup V_2$,

$$\chi_i^{\text{val}}(n) = \begin{cases} \chi_i^{\text{val}}(n) & n \in N_1 \\ \chi_i^{\text{val}}(n') & n = n' \oplus s, n' \in N_2 \end{cases} \quad \chi^{\text{type}}(n) = \begin{cases} \chi^{\text{type}}(n) & n \in N_1 \\ \chi^{\text{type}}(n') & n = n' \oplus s, n' \in N_2 \end{cases}$$

- *Sorting function*: let $D = (I, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})$ be a dataframe and $n \in N$ an attribute name. Let $b_n : I \rightarrow \{1, \dots, |I|\}$ a bijection such that:
 - If $\chi_i^{\text{val}}(n) \leq_V \chi_j^{\text{val}}(n)$ according to the partial order defined on V , then $b_n(i) \leq_{\mathbb{N}} b_n(j)$.
 - If $\chi_i^{\text{val}}(n) \leq_V \chi_j^{\text{val}}(n) \wedge \chi_j^{\text{val}}(n) \leq_V \chi_i^{\text{val}}(n)$, then $b_n(i) \leq_{\mathbb{N}} b_n(j) \iff i \leq_{\mathbb{N}} j$.


With this definition, $\text{sort}(D, n, b) = (\{1, \dots, |I|\}, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})$ could be introduced as $\chi_{b_n(i)}^{\text{val}}(n) = \chi_i^{\text{val}}(n)$.

- *String attribute merging*: let $D = (I, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})$ be a dataframe, then a new attribute $n \in \mathcal{U}_{\text{attr}}$ of type $\text{str} \in \mathcal{U}_T$ could be inserted into the dataframe as merge of the values of two (existing) attributes $n_1, n_2 \in N$ (separated by a string delimiter s) to obtain a new dataframe


$$\text{mergstrv}(D, n, n_1, n_2, s) = (I, N', T, V', \chi_i^{\text{val}}, \chi^{\text{type}})$$

where: $N' = N \cup \{n\}$ and $V' = V \cup \cup_{i \in I} \{\chi_i^{\text{val}}(n_1) \oplus s \oplus \chi_i^{\text{val}}(n_2)\}$

	case:concept:name	concept:name
	1A	1A
	1A	1B
	1B	1C
	1C	2A
	2A	2B
	2B	2C
	2C	



Sorted dataframe



Shifted sorted dataframe

Fig. 3. Example of the calculation of the Directly-Follows Graph on top of dataframes, using the shifting and counting method.

5.4 Possibility to calculate the Directly-Follows Graph on top of dataframes

In this section, we want to show that it's actually possible to calculate the Directly-Follows Graph (DFG; see subsection 3.2) on top of the dataframe structures. Let $D = (I, N, T, V, \chi_i^{\text{val}}, \chi^{\text{type}})$ be a dataframe. Let $n_{\text{case}} \in N$ be the case identifier attribute, and $n_{\text{act}} \in N$ be the attribute that identifies the activity. The nodes of the DFG are simply the different values for the activities: $\pi_2(\chi^{\text{val}}(n_{\text{act}}))$. For the retrieval of the edges and the count, two different strategies are actually possible:

- *Map-reduce*: the dataframe is grouped using the n_{case} attribute, and then the edges and the counts are retrieved as:

$$E = \bigcup_{G \in \text{group}(D, n_{\text{case}})} \bigcup_{(i,j) \in \pi_I(G)^2, i < j \wedge \exists k \in \pi_I(G), i < k < j} \{(\chi_i^{\text{val}}(n_{\text{act}}), \chi_j^{\text{val}}(n_{\text{act}}))\}$$

For $e \in E$,

$$c(e) = \sum_{G \in \text{group}(D, n_{\text{case}})} \sum_{(i,j) \in \pi_I(G)^2, i < j \wedge \exists k \in \pi_I(G), i < k < j} \mathbb{1}_{(\chi_i^{\text{val}}(n_{\text{act}}), \chi_j^{\text{val}}(n_{\text{act}})) = e}$$

In a map-reduce implementation, the grouping is performed and then different groups can be assigned to different workers (*map* phase). The single results are then merged in the *reduce* phase to obtain the final result.

- *Shifting and counting*: the strategy assumes that the dataframe is sorted the strategy could be explained as follows:

1. A shifted version of the dataframe is obtained using the shift function.
2. Then, a concatenated dataframe is obtained as: $D' = (I', N', T', V', \chi_i^{\text{val}}, \chi^{\text{type}}) = \text{concat}(D, \text{shift}(D), "2")$.
3. Then, if $n_{\text{case}}, n'_{\text{case}}$ are the attributes identifying the case respectively in D and D' , a dataframe D'' could be obtained taking $S = \{n_{\text{case}}, n'_{\text{case}}\}$, and $f : I \times \mathcal{P}(N) \rightarrow \{0, 1\}$ such that $f(i, S) = \mathbb{1}_{\chi_i^{\text{val}}(n_{\text{case}}) = \chi_i^{\text{val}}(n'_{\text{case}})}$, as $D'' = (I'', N', T', V', \chi_i^{\text{val}}, \chi^{\text{type}}) = \text{proj}(D', S, f)$. In such way, only rows in the concatenated dataframe that have the same case ID are kept.

- Eventually, a new dataframe with an additional column n''_{act} can be defined starting from D'' , if n_{act}, n'_{act} are the attributes identifying the activity respectively in D and D' , as:

$$D''' = (I''', N'', T'', V'', \chi_i'''^{val}, \chi_i'''^{type}) = \text{mergstrv}(D'', n''_{act}, n_{act}, n'_{act}, ", ")$$

- The edges and the count of the DFG are: $E = \pi_2(\chi_i'''^{val}(n''_{act}))$ and $c : E \rightarrow \mathbb{N}, c(e) = |\{i \in I | \chi_i'''^{val}(n''_{act}) = e\}|$

A representation of the ideas underlying the algorithm is contained in Figure 3.

5.5 Complexity Estimation on Logs and Dataframe Implementations

Filtering: Filtering on the attribute values of events is an ubiquitous operation in process mining, since only a part of the events of the log may be of interest for an analysis. Even in the worst case, accessing the value of an attribute for an event has $O(1)$ complexity, hence filtering a dataframe with N events has a complexity that is in the worst case equal to $O(N)$. A comparison between the complexity of filtering on event logs and dataframe implementations is contained in subsection 5.5.

Table 3. Comparison between the theoretical complexity of filtering on the attribute values of events on implementation of logs and dataframe structures (average and worst case). Here, N is the number of events in the log, and M is the number of different names for the attributes. Since the access to attribute values on classic event logs require access to a map, the complexity of filtering in the worst case is greater than the complexity of filtering operations on dataframes.

Implementation	Complexity - Average	Complexity - Worst
Iteration on a classic event log	$O(N)$	$O(N * M)$
Dataframe	$O(N)$	$O(N)$

Directly-Follows Graph computation: the computation of the DFG on top of classical event logs requires a single iteration on the log and the storage of the DFG edges in a map. Hence, the complexity is $O(N)$ in the average case and $O(N^2)$ in the worst case where all the events belong to the same case and a lot of collisions happen in the map. As seen in subsection 5.4, there are two ways to calculate the DFG on top of dataframes: map-reduce, and “shifting and counting”.

For the map-reduce approach, the complexity is still $O(N)$ in the average case because all the operations involved (grouping, mapping and reducing to the map of edges and counts) have a linear average complexity. The reduce approach, in the worst case, can have $O(N^2)$ complexity when a lot of collisions inside the map happen. However, the real-life map-reduce implementations on dataframes (Pandas, Spark) are disappointing, being slow in the map operation and/or hampered by network traffic issues (when the computations are split between different nodes)

For the “shifting and counting” approach, the initial sort on the dataframe has a complexity (if algorithms like mergesort and heapsort are used) equal to $O(N * \log(N))$ in the average and worst case. Then, every remaining operation has linear complexity, hence excluding the sorting operation, the complexity of the remaining operations is still $O(N)$ in the average case, and $O(N^2)$ in the worst (that happens when a lot of collisions happen in writing the edges and their count in the map).

To recap, the main difference between classical event logs and dataframes is the way attributes at event level are accessed:

- On classical event logs, the attribute map of the event is retrieved, and a specific key of the map is accessed to retrieve the value of the attribute.
- On dataframes, the list of values for the specific attribute is retrieved, and then a specific index (corresponding to the index of the row/event) is retrieved.

A comparison between the filtering complexity on classical event logs and dataframe implementations is contained in Table 5.5.

Table 4. Comparison between the theoretical complexity of DFG calculation on implementations of logs and dataframe structures (average and worst case). Here, N is the number of events in the log. The iteration, the map-reduce and the shifting approach (although, this one assumes that a sorting of the dataframe on the case identifier is done) share the same average and worst case complexity.

Implementation	Complexity - Average	Complexity - Worst
Iteration on a classic event log	$O(N)$	$O(N^2)$
Dataframe - Map-Reduce	$O(N)$	$O(N^2)$
Dataframe - Shifting(sort by Case ID to be done)	$O(N * \log(N))$	$O(N^2)$
Dataframe - Shifting (sort by Case ID already done)	$O(N)$	$O(N^2)$

6 Assessment

In this section, an assessment of the speed of a dataframe implementation (Pandas), that can ingest data from the Parquet columnar format is done on some popular event logs (a list of logs and features is contained in section 1). While a direct comparison is avoided (it would be probably unfair to compare the ProM6 process mining framework, the final goal is to show that implementations of the dataframe structure can handle well modern real-life event logs, using the advantages provided by columnar log formats (higher compression ratios, possibility to choose the attributes at load time). The results are obtained using the PM4Py python process mining library (<http://www.pm4py.org>), importing/exporting dataframes from/to Parquet files (see <http://pm4py.pads.rwth-aachen.de/documentation/working-with-event-data/>) and using dataframe computation techniques offered by the library (see <http://pm4py.pads.rwth-aachen.de>).

de/documentation/big-dataframe-management/). The methodology and the results of the assessment are contained in tables 5 and 6. Some further evidence that a dataframe structure provides a significant support for process mining operations is contained in the PM4Py web services[7]. There, different handlers have been built for classic event logs (ingested from XES) and dataframes (ingested from CSV or Parquet).

Table 5. Assessment of columnar storages and dataframes on real-life logs (available in the 4TU repository). All the experiments (except the DFG calculation) are repeated 1) including all the attributes, 2) including only the case ID and the activity column. The experiments are: the measurement of the size on disk (MB), the loading time (in seconds), the amount of RAM needed to store the Pandas dataframe, the time (in seconds) needed to filter the dataframe keeping events of the most common activity, and the DFG calculation time (in seconds) using the shifting and counting algorithm. The measurements have been done on a Lenovo ThinkPad T470s notebook (I7-7550U, 16 GB RAM, DDR) and show that columnar storages/dataframes can handle very well current real-life logs up to 2.5 millions of events.

Log	S.Parq.	2c	LoadT.	2c	RAM	2c	A.Filt.	2c	DFG(2c)
roadtraffic	6MB	4MB	0.42s	0.18s	73MB	13MB	0.30s	0.12s	0.24s
bpic2017_o	7MB	1.7MB	0.32s	0.13s	24MB	4MB	0.11s	0.04s	0.09s
bpic2017_a	29MB	5.7MB	1.93s	0.33s	184MB	28MB	0.60s	0.16s	0.59s
bpic2018	167MB	11.8MB	4.88s	0.43s	786MB	58MB	2.68s	0.36s	1.17s
bpic2019	14MB	3.1MB	2.92s	0.59s	234MB	24MB	1.37s	0.37s	0.88s

Table 6. Assessment of columnar storages and dataframes on synthetic big logs (available at the address <https://drive.google.com/open?id=1VgcELM6NqJxWxs3kC2j0ZbYpK2F1w9tV>). The experiments are: the measurement of the size on disk (MB), the loading time (in seconds), the amount of RAM needed to store the Pandas dataframe, the time (in seconds) needed to filter the dataframe keeping events of the most common activity, and the DFG calculation time (in seconds) using the shifting and counting algorithm. The measurements have been done on a Lenovo ThinkPad T470s notebook (I7-7550U, 16 GB RAM, DDR) and show that dataframes can efficiently handle big logs of 10^6 cases and 10^7 events.

Log	N.Cas.	N.Ev.	Disk(MB)	Load	RAM(MB)	Filt.	DFG
L1	1000000	7003896	78	2.60s	300	0.73s	4.39s
L2	2000000	14001120	157	4.65s	650	1.36s	8.85s
L3	3000000	20999048	237	7.33s	950	2.00s	13.17s
L4	4000000	34989340	396	12.66s	1600	3.38s	22.49s
L5	5000000	48988692	554	17.08s	2200	4.99s	33.48s

7 Conclusion

In this paper, a combination of columnar storages (like Apache Parquet and Apache ORC) and dataframes (popular implementations are Pandas and Apache Spark) is used in order to increase the scalability of process mining techniques.

All the described techniques are implemented in the PM4Py library[8] and the PM4Py web services[7]. The advantages of existing columnar formats are the efficient compression techniques that are available when data of the same attribute (having always the same type) is saved in the same record, and the attribute selection at load time (that helps to avoid loading unnecessary attributes). For dataframes, while some popular implementations already exist, this paper (at the best of the authors knowledge) provides the first abstraction of the concept of dataframe, describing a set of operations that are feasible on the data structure and how to obtain a conversion from classical event logs to a dataframe structure. The provision of such structure, and of a basic implementation of dataframe, is important to understand the differences with the classical event log and to provide a set of algorithms on top of dataframes. Theoretical complexity is evaluated for two mainstream process mining techniques, for both algorithms on classical event logs implementations and the basic dataframe implementation.

The theoretical complexity estimation and the assessment shows that dataframes, combined with columnar storages, can provide a way to efficiently handle millions of events. This is thanks to specific algorithms that are defined on the dataframe structure (filtering at the event level, and DFG calculation). Hence, the techniques described in this paper seem suited to increase process mining scalability. Direct comparison with the event log structure is avoided since the underlying frameworks/languages are different (ProM6/Java and PM4Py/Python), hence fairness of the analysis would be at risk. However, to make use of the dataframe structure, specific algorithms are needed. This requires a complete re-engineering of the process mining techniques. For example, filtering at case level requires custom techniques that could be less efficient than the ones available for classical event logs. Authors have already taken into PM4Py several dataframe-specific techniques (filtering at event and case level, statistics for cases/variants, LTL Checker). Another limitation is the fact that, while in the XES standard, some metadata and additional information related to the process is stored inside the XML file, the possibility to store them in columnar storages like Apache Parquet and Apache ORC has not been cleared in the pages of this paper.

XES is, and will be, an IEEE standard format that ensures interoperability between different process mining tools. It should also be said that Parquet is a format that in the latest years is very used in data extraction from information systems, hence having an efficient structure on top of columns improves the integration with databases and processing techniques. Overall, this paper is an introduction to columnar storages and dataframes in the field of process mining. Some further development should be expected on the data partitioning for sampling and the distributed application of process mining algorithms.

References

1. van der Aalst, W., Damiani, E.: Processes meet big data: Connecting data science with process science. *IEEE Transactions on Services Computing* **8**(6) (2015)

2. Abadi, D., Boncz, P., Harizopoulos, S., Idreos, S., Madden, S., et al.: The design and implementation of modern column-oriented database systems. *Foundations and Trends® in Databases* **5**(3), 197–280 (2013)
3. Abadi, D.J., Madden, S.R., Hachem, N.: Column-stores vs. row-stores: how different are they really? In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. pp. 967–980. ACM (2008)
4. Adriansyah, A., van Dongen, B., van der Aalst, W.: Conformance checking using cost-based fitness analysis. In: *2011 IEEE 15th International Enterprise Distributed Object Computing Conference*. pp. 55–64. IEEE (2011)
5. Augusto, A., Conforti, R., Dumas, M., La Rosa, M., Maggi, F.M., Marrella, A., Mecella, M., Soo, A.: Automated discovery of process models from event logs: Review and benchmark. *IEEE Transactions on Knowledge and Data Engineering* **31**(4), 686–705 (2018)
6. Azzini, A., Ceravolo, P.: Consistent process mining over big data triple stores. In: *2013 IEEE International Congress on Big Data*. pp. 54–61. IEEE (2013)
7. Berti, A., van Zelst, S.J., van der Aalst, W.: PM4Py Web Services: Easy Development, Integration and Deployment of Process Mining Features in any Application Stack. In: *BPM Demo Track* (2019)
8. Berti, A., van Zelst, S.J., van der Aalst, W.: Process Mining for Python (PM4Py): Bridging the Gap Between Process-and Data Science. In: *ICPM Demo Track (CEUR 2374)*. pp. 13–16 (2019)
9. Cheng, L., Van Dongen, B., Van Der Aalst, W.: Scalable Discovery of Hybrid Process Models in a Cloud Computing Environment. *IEEE Transactions on Services Computing* (2019)
10. van Dongen, B.F., Shabani, S.: Relational XES: Data Management for Process Mining. In: *CAiSE Forum*. pp. 169–176 (2015)
11. Hernández, S., van Zelst, S.J., Ezpeleta, J., van der Aalst, W.: Handling Big (ger) Logs: Connecting ProM 6 to Apache Hadoop. In: *BPM (Demos)*. pp. 80–84 (2015)
12. Leemans, S.J., Fahland, D., van der Aalst, W.: Discovering block-structured process models from event logs—a constructive approach. In: *International conference on applications and theory of Petri nets and concurrency*. Springer (2013)
13. Leemans, S.J., Fahland, D., van der Aalst, W.: Scalable process discovery and conformance checking. *Software & Systems Modeling* **17**(2), 599–631 (2018)
14. Leemans, S.J., Poppe, E., Wynn, M.T.: Directly follows-based process mining: Exploration & a case study. In: *Proceedings of the ICPM International Conference on Process Mining, Aachen, Germany*. p. 46 (2019)
15. Loreti, D., Chesani, F., Ciampolini, A., Mello, P.: Distributed compliance monitoring of business processes over mapreduce architectures. In: *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion*. pp. 79–84. ACM (2017)
16. Mannhardt, F.: XESLite-managing large XES event logs in ProM. *BPM Center Report BPM-16-04* pp. 224–236 (2016)
17. Syamsiyah, A., van Dongen, B., van der Aalst, W.M.: Db-xes: enabling process discovery in the large. In: *International Symposium on Data-Driven Process Discovery and Analysis*. pp. 53–77. Springer (2016)
18. Valdivieso, H., Lee, W.L.J., Munoz-Gama, J., Sepúlveda, M.: OpyenXES: A Complete Python Library for the eXtensible Event Stream Standard pp. 71–75 (2018)
19. Verbeek, H., Buijs, J.C., Van Dongen, B., van der Aalst, W.: Xes, xesame, and prom 6. In: *International Conference on Advanced Information Systems Engineering*. pp. 60–75. Springer (2010)