iProcess: Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes

Amin Beheshti[1(✉)], Francesco Schiliro[1,2], Samira Ghodratnama[1], Farhad Amouzgar[1], Boualem Benatallah[3], Jian Yang[1], Quan Z. Sheng[1], Fabio Casati[4], and Hamid Reza Motahari-Nezhad[5]

[1] Macquarie University, Sydney, Australia
{amin.beheshti,jian.yang,michael.sheng}@mq.edu.au,
{francesco.schiliro,samira.ghodratnama,farhad.amouzgar}@hdr.mq.edu.au
[2] Australia Federal Police, Canberra, Australia
[3] University of New South Wales, Sydney, Australia
boualem@cse.unsw.edu.au
[4] University of Trento, Trento, Italy
fabio.casati@unitn.it
[5] EY AI Lab, Palo Alto, USA
hamid.motahari@ey.com

Abstract. The Internet of Things (IoT), the network of physical objects augmented with Internet-enabled computing devices to enable those objects sense the real world, has the potential to transform many industries. This includes harnessing real-time intelligence to improve risk-based decision making and supporting adaptive processes from core to edge. For example, modern police investigation processes are often extremely complex, data-driven and knowledge-intensive. In such processes, it is not sufficient to focus on data storage and data analysis; and the knowledge workers (e.g., investigators) will need to collect, understand and relate the big data (scattered across various systems) to process analysis: in order to communicate analysis findings, supporting evidences and to make decisions. In this paper, we present a scalable and extensible IoT-Enabled Process Data Analytics Pipeline (namely *iProcess*) to enable analysts ingest data from IoT devices, extract knowledge from this data and link them to process (execution) data. We introduce the notion of process *Knowledge Lake* and present novel techniques to summarize the linked IoT and process data to construct process *narratives*. This enables us to put the first step towards enabling *storytelling* with process data.

Keywords: Process data science · Process Data Analytics
Data-driven business processes · Knowledge-intensive business processes

M. Weske et al. (Eds.): BPM Forum 2018, LNBIP 329, pp. 108–126, 2018.
https://doi.org/10.1007/978-3-319-98651-7_7

1 Introduction

Information processing using knowledge-, service-, and cloud-based systems has become the foundation of the twenty-first-century life. Recently, the focus of process thinking has shifted towards understanding and analyzing process related data captured in various information systems and services that support processes [2,7,8]. The Internet of Things (IoT), i.e., the network of physical objects augmented with Internet-enabled computing devices to enable those objects sense the real world, has the potential to generate large amount of process related data which can transform many industries. This includes harnessing real-time intelligence to improve risk-based decision making and supporting adaptive processes from core to edge. For example, modern police investigation processes are often extremely complex, data-driven and knowledge-intensive. Considering cases such as Boston bombing (USA), the ingestion, curation and analysis of the big data generated from various IoT devices (CCTVs, Police cars, camera on officers on duty and more) could be vital but is not enough: the big IoT data should be linked to process execution data and also need to be related to process analysis. This will enable organizations to communicate analysis findings, supporting evidences and to make decisions.

Current state-of-the-art in analyzing business processes does not provide sufficient data-driven techniques to relate IoT and process related data to process analysis and to improve risk-based decision making in knowledge intensive processes. To address this challenge, in this paper, we present a scalable and extensible IoT-Enabled Process Data Analytics Pipeline to enable analysts to ingest data from IoT devices, extract knowledge from this data and link them to process (execution) data. We present novel techniques to summarize the linked IoT and process data to construct *process narratives*. Finally, we offer a Machine-Learning-as-a-Service layer to enable process analysts to analyze the narratives and dig for facts in an easy way. We adopt a motivating scenario in policing, where a knowledge worker (e.g., a criminal investigator) in a knowledge intensive process (e.g., criminal investigation) will be augmented by smart devices to collect data on the scene as well as locating IoT devices around the investigation location and communicate with them to understand and analyze evidences in real time. This paper includes offering:

- A scalable and extensible IoT-Enabled Process Data Analytics Pipeline to enable analysts to ingest data from IoT devices, extract knowledge from this data and link them to process (execution) data. We leverage our previous work [3,9] to ingest and organize the big IoT and process data in Data Lakes [3] and to automatically contextualize the raw data in the Data Lake and construct a Knowledge Lake [4].
- A framework and algorithms for *summarizing* the (big) process data and constructing process narrative. We present a set of innovative, fine-grained and intuitive analytical services to discover patterns and related entities, and enrich them with complex data structures (e.g., timeseries, hierarchies and subgraphs) to construct *narratives*.

– A spreadsheet-like dashboard to enable analysts interact with narratives and control their resolution in an easy way. We present a machine-learning-as-a-service framework, which enable analysts dig for facts in an easy way.

The rest of the paper is organized as follows. In Sect. 2 we provide the related work and a motivating scenario. We present the IoT-Enabled Process Data Analytics Pipeline in Sect. 3. We discuss the implementation and the evaluation in Sect. 4 before concluding the paper in Sect. 5.

2 Related Work and Motivating Scenario

2.1 Internet of Things

The Internet of Things (IoT) has the potential to transform many industries and enable them to harness real-time intelligence to improve risk-based decision making and to support adaptive processes from core to edge. In IoT, many of the objects that surround us will be connected, and will be sensing the real world. These objects have the potential to generate large amount of data and meta-data which may contain various facts and evidences. These facts and evidences can help knowledge workers understand knowledge intensive processes and make correct decisions [19]. Many of the work in IoT focus on applications such as smart and connected communities [22], industries (e.g., agriculture, food processing, environmental monitoring, automotive, telecommunications, and health) [15], and security and privacy [1]. Mobile crowdsensing and cyber-physical cloud computing presented as two most important IoT technologies in promoting Smart and Connected Communities [22]. Management of IoT data is an important issue in rapidly changing organizations. A set of recent work has been focusing on ingesting the large amount of data generated from IoT devices and store and organize them in big data platforms. For example, Hortonworks DataFlow (hortonworks.com) provides an end-to-end platform that collects and organizes the IoT data in the cloud. Other approaches include Teradata (teradata.com/) and Oracle BigData (oracle.com/bigdata) focus on data management and analytics, and do not related the data to process analysis.

Enabling IoT data in business process analytics, as presented in this paper, is a novel approach to enhance data-driven techniques for improving risk-based decision making in knowledge intensive processes. The novel notions of Knowledge Lake and narrative, presented in this paper, will enable us to put the first step towards enabling *storytelling* with process data. This will enable analysts to ingest data from IoT devices, extract knowledge from this data and related the data to process analysis.

2.2 Data-Driven Processes

The problem of understanding the behavior of information systems as well as the processes and services they support has become a priority in medium and large enterprises. This is demonstrated by the proliferation of tools for the analysis of

process executions, system interactions, and system dependencies, and by recent research work in process data warehousing, discovery and mining [24]. Accordingly, identifying business needs and determining solutions to business problems requires the analysis of business process data which in turn will help in discovering useful information and supporting decision making for enterprises. The state-of-the-art in process data analytics focus on various topics such as Warehousing Business Process Data [14], Data Services and DataSpaces [13], Supporting Big Data Analytics Over Process Execution Data [5], Process Spaces [18], Process Mining [24] and Analyzing Cross-cutting Aspects (e.g., provenance) in Processes' Data [6]. In our recent book [8], we provided a complete state-of-the-art in the area of business process management in general and process data analytics in particular. This book provides defrayals on: (i) technologies, applications and practices used to provide process analytics from querying to analyzing process data; (ii) a wide spectrum of business process paradigms that have been presented in the literature from structured to unstructured processes; (iii) the state-of-the-art technologies and the concepts, abstractions and methods in structured and unstructured BPM including activity-based, rule-based, artifact-based, and case-based processes; and (iv) the emerging trend in the business process management area such as: process spaces, big-data for processes, crowdsourcing, social BPM, and process management on the cloud.

Summarization techniques presented in this paper, is a novel approach to enable analysts to understand and relate the big IoT and process data to process analysis in order to communicate analysis findings and supporting evidences in an easy way. The proposed approach will enhance data-driven techniques for improving risk-based decision making in knowledge intensive processes.

2.3 Knowledge-Intensive Processes

Case-managed processes are primarily referred to as semistructured processes, since they often require the ongoing intervention of skilled and knowledgeable workers. Such Knowledge-Intensive Processes, involve operations that heavily reliant on professional knowledge. For these reasons, it is considered that human knowledge workers are responsible to drive the process, which cannot otherwise be automated as in workflow systems [8]. Knowledge-intensive processes almost always involve the collection and presentation of a diverse set of artifacts and capturing the human activities around artifacts. This, emphasizes the artifact-centric nature of such processes. Many approaches [11,16,23] used business artifacts that combine data and process in a holistic manner and as the basic building block. Some of these works [16] used a variant of finite state machines to specify lifecycles. Some theoretical works [11] explored declarative approaches to specifying the artifact lifecycles following an event oriented style. Another line of work in this category, focused on querying artifact-centric processes [17].

Another related line of work is artifact-centric workflows [11] where the process model is defined in terms of the lifecycle of the documents. Some other works [20,21], focused on modeling and querying techniques for knowledge-intensive tasks. Some of existing approaches [20] for modeling ad-hoc processes

focused on supporting ad-hoc workflows through user guidance. Some other approaches [21] focused on intelligent user assistance to guide end users during ad-hoc process execution by giving recommendations on possible next steps. Another line of work [6], considers entities (e.g., actors, activities and artifacts) as first class citizens and focuses on the evolution of business artifacts over time. Unlike these approaches, in *iProcess*, we not only consider artifacts as first class citizens, but we take the information-items (e.g., named entities, keywords, etc.) extracted from the content of the artifacts into account.

2.4 Motivating Scenario: Missing People

As the motivating scenario, we focus on the investigation processes around *Missing Persons*. Between 2008 and 2015 over 305,000 people were reported missing in Australia (aic.gov.au/), an average of 38,159 reports each year. In USA (nij.gov/), on any given day, there are as many as 100,000 active missing person's cases. The first few hours following a person's disappearance are the most crucial. The sooner police is able to put together the sequence of events and actions right before the disappearance of the person, the higher the chance of finding the person. This entails gathering information about the person including physical appearance, and activities on social media in the physical/social environments of the person, person's activity data such as phone calls and emails, and information on the person detected by sensors (e.g. CCTVs).

The investigation process is a data-driven, knowledge-intensive and collaborative process. The information associated with an investigation (case process) are usually complex, entailing the collection and presentation of many different types of documents and records. It is also common that separate investigations may impact other investigation processes, and the more evidences (knowledge and facts extracted from the data in the data lake [3]) collected the better related cases can be linked explicitly. Although law enforcement agencies use data analysis, crime prevention, surveillance, communication, and data sharing technologies to improve their operations and performance, in sophisticated and data intensive cases such as missing persons there still remain many challenges. For example, fast and accurate information collection and analysis is vital in law enforcement applications [10,12]. From the policymakers' perspective, this trend calls for the adoption of innovations and technologically advanced business processes that can help law enforcers detect and prevent criminal acts. Enabling IoT data in law enforcement processes will help investigators to access to a potential pool of data evidences. Then, the challenge would be to prepare the big process data for analytics, summarizing the big process data, constructing narratives and enable analysts to link narratives and dig for facts in an easy way.

In this paper, we aim to address this challenge by augmenting police officers with Internet-enabled smart devices (e.g., phones/watches) to assist them in the process of collecting evidences, access to location-based services to identify and locate resources (CCTVs, camera on officers on duty, police cars, drones and more), organize all these islands of data in a Knowledge Lake [4] and feed them into a scalable and extensible IoT-Enabled Process Data Analytics Pipeline.

3 iProcess: IoT-Enabled Process Data Analytics Pipeline

Figure 1 illustrates the IoT-Enabled Process Data Analytics Pipeline framework. In the following we explain the main phases of the iProcess pipeline.

3.1 Process Data-Lake

In order to understand data-driven knowledge-intensive processes, one may need to perform considerable analytics over large hybrid collections of heterogeneous and partially unstructured data that is captured from private (personal/business), social and open data. Enabling IoT data in such processes will maximize the value of data-in-motion and will require dealing with big data organization challenges such as wide physical distribution, diversity of formats, non-standard data models, independently-managed and heterogeneous semantics. In such an environment, analysts may need to deal with a collection of datasets, from relational to NoSQL, that holds a vast amount of data gathered from various data islands, i.e., Data Lake. To address this challenge, we leverage our previous work [3], CoreDB: a Data Lake as a Service, to identify (IoT, Private, Social and Open) data sources and ingest the big process data in the Data Lake. CoreDB manages multiple database technologies (from relational to NoSQL), offers a built-in design for security and tracing, and provides a single REST API to organize, index and query the data and metadata in the Data Lake.

3.2 Process Knowledge-Lake

The rationale behind a Data Lake is to store raw data and let the data analyst decide how to cook/curate them later. We introduce the notion of Knowledge Lake [4], i.e., a contextualized Data Lake, to provide the foundation for big data analytics by automatically curating the raw data in the Data Lake and to prepare them for deriving insights. To achieve this goal, we leverage our previous work [4], to transform raw data (unstructured, semi-structured and structured data sources) into a contextualized data and knowledge that is maintained and made available for use by end-users and applications. The Data Curation APIs [9] in the Knowledge Lake provide curation tasks such as extraction, linking, summarization, annotation, enrichment, classification and more. This will enable us to add features - such as extracting keyword, part of speech, and named entities such as Persons, Locations and Organizations; providing synonyms and stems for extracted information items leveraging lexical knowledge bases for the English language such as WordNet; linking extracted entities to external knowledge bases such as Google Knowledge Graph and Wikidata; discovering similarity among the extracted information items; classifying, indexing, sorting and categorizing data - into the data and knowledge persisted in the Knowledge Lake.

This will enable us, for example, to extract and link information about the missing person from various data islands in the data lake such as the IoT, social and news data sources and to relate them to missing person case. The goal of this

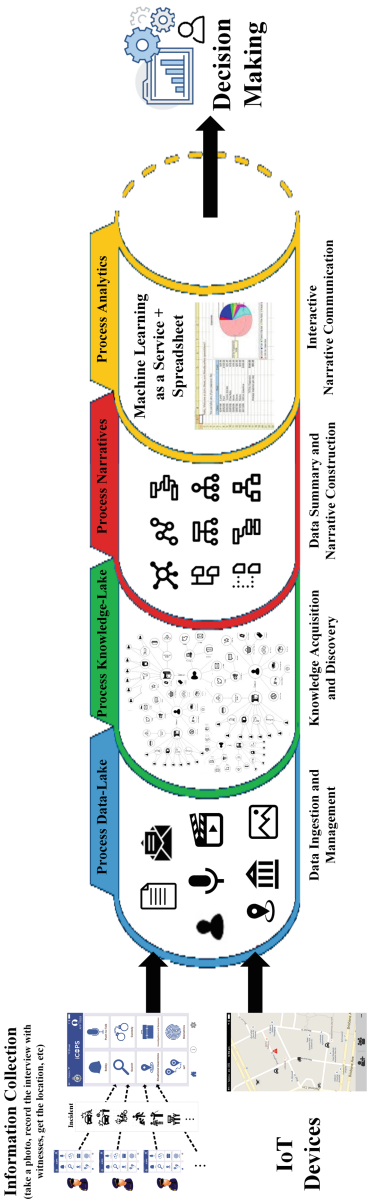

Fig. 1. IoT-Enabled process data analytics pipeline.

phase is to contextualize the Data Lake and turn it into a Process Knowledge-Lake which contains: (i) a set of facts, information, and insights extracted from the raw data; (ii) process event data i.e., observed behavior; and (iii) process models, e.g., manually or automatically discovered. All these three main components will enable the process analysts to relate data to process analysis. To achieve this goal, we present a graph model to define the entities (process data, instances and models) and the relationships among them.

Definition 1 (Process Knowledge Graph)**.** Let $G = (V, E)$ be an Entity-Relationship (ER) attributed graph where V is a set of nodes with $|V| = n$, and $E \subseteq (V \times V)$ is a set of ordered pairs called edges. Let $H = (V, E)$ be a RDF graph where V is a set of nodes with $|V| = n$, and $E \subseteq (V \times V)$ is a set of ordered pairs called edges. An ER graph $G = (V_G, E_G)$ with n entities is defined as $G \subseteq H$, $V_G = V$ and $E_G \subseteq E$ such that G is a directed graph with no directed cycles. We define a resource in an ER graph recursively as follows: (i) The sets V_G and E_G are resources; (ii) $\in$ is a resource; and (iii) The set of ER graphs are closed under intersection, union and set difference: let G_1 and G_2 be two ER graphs, then $G_1 \cup G_2$, $G_1 \cap G_2$, and $G_1 - G_2$ are resources.

Definition 2 (Entity)**.** An entity E is represented as a data object that exists separately and has a unique identity. Entities are described by a set of *attributes* but may not conform to an entity type. Entities can be complex such as Process Model, Process Instance and a (IoT, Social or private) Data Source. One way would be to define "stream events" meaning events that are tied to a specific timestamp or sequence number, and associated to a specific IoT device. Entities can be also simple such as *artifacts* (e.g., structured such as customer record or unstructured such as an email), actors and activities. Entities can be atomic *information items* such as a keyword, phrase, topic and named entity (e.g., people, location, organization) extracted from unstructured artifacts such as emails, images (extracted from IoT devices) or social items (such as a Tweet in Twitter). This entity model offers flexibility when types are unknown and takes advantage of structure when types are known. Entities can be of type stream, such as 'stream events' meaning events that are tied to a specific timestamp or sequence number, and associated to a specific IoT device.

Definition 3 (Relationship)**.** A *relationship* is a directed link between a pair of entities, which is associated with a predicate defined on the attributes of entities that characterizes the relationship. Relationships can be described by a set of *attributes* but may not conform to a relationship type. Relationships can be [2]: Time-based, Content-based and Activity-based. We define the following *explicit* relationships:

- $Process \xrightarrow{(Instance\text{-}of)} Model$: express that a process is an instance of a process model.
- $Process \xrightarrow{(Used)} Artifact$: express that a process used an artifact during its execution.

- $Artifact \xrightarrow{(Generated\text{-}by)} Process$: express that an artifact was generated by a process.
- $Process \xrightarrow{(Controlled\text{-}by\ (R))} Actor$: express that a process was controlled by an actor. Given that a process may have been controlled by several actors, it is important to identify the roles of actors.
- $Process_1 \xrightarrow{(Triggered\text{-}by)} Process_2$: express a process oriented view where a process triggered another process.
- $Artifact \xrightarrow{(Organized\text{-}in)} Data-Island$: express that an artifact (e.g., an email in a private dataset or an image extracted from a CCTV camera) is organized in a Data Island (i.e., a Data source in the Data Lake).
- $Information-Item \xrightarrow{(Extracted\text{-}from)} Artifact$: express that an information item (e.g., a topic extracted from a Tweet or a named entity such as a person, extracted from an Image) is extracted from an artifact (e.g., an email or an image, extracted from a CCTV camera, in a private data source).
- $Information-Item_1 \xrightarrow{(Similar\text{-}to)} Information-Item_2$: express that an information item (e.g., a person named entity extracted from an Image) is similar to another information item (e.g., a person named entity extracted from an email or a Tweet in Twitter (twitter.com)

Notice that 'Process' refers to a process instance and 'Model' refers to a process model. A *Process Instance or Case*, is a triple $C = (P_F, N_{start}, N_{end})$, where P_F is a path in which the nodes in P are of type 'event', grouped using the function F (e.g. a function can be a 'Correlation Condition'), and are in chronological order. A *Process Model*, allows the generation of all valid (acceptable) case C of a process, e.g. implemented by service or a set of services [2]. Various process mining algorithms and tools (e.g., PROM), include our previous work [18], can be used to automatically extract the first type of relationship. Process instances and services can be instrumented to automatically construct the other type of relationship.

3.3 Process Narratives

In this phase, we present an *OLAP* [5] *style process data summarization* technique as an alternative to querying and analysis techniques. This approach will isolate the process analyst from the process of explicitly analyzing different dimensions such as time, location, activity, actor and more. Instead, the system will be able to use interactive (artifacts, actors, events, tasks, time, location, etc.) summary generation to select and sequence narratives dynamically. This novel summarization method will enable process analysts to choose one or more dimensions (i.e., attributes and relationships), based on their specific goal, and interact with small and informative summaries. This will enable the process analysts to analyze the process from various dimensions. Figure 2(B) illustrates a sample OLAP dimension.

In OLAP [5], cubes are defined as set of partitions, organized to provide a multi-dimensional and multi-level view, where partitions considered as the unit

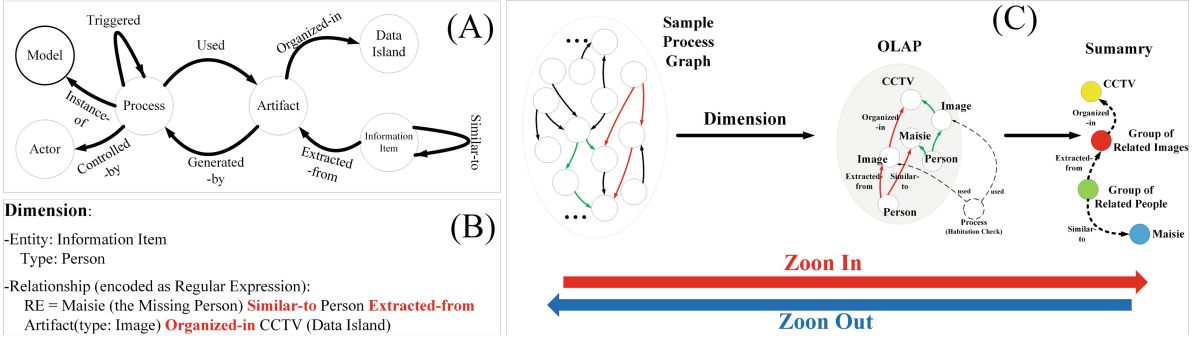

Fig. 2. Process Knowledge Graph schema (A), a sample OLAP Dimension (B) and an interactive graph summary (C).

of granularity. Dimensions defined as perspectives used for looking at the data within constructed partitions. In police investigation scenarios, such as Boston bombing, process cubes can enable effective analysis of the Process Knowledge Graph from different perspectives and with multiple granularities. For example, by aggregating and relating all evidences from the person of interest, location of the incident and more. Following, we define a process cube.

Definition 4 (Process Cube)**.** A process cube defined to extend decision support on multidimensional networks, e.g., process graphs, considering both data objects and the relationships among them. We reuse and extend the definition for graph-cube proposed in our previous work [5]. In particular, given a multidimensional network N, the graph cube is obtained by restructuring N in all possible aggregations of set of node/edge attributes A, where for each aggregation A' of A, the measure is an aggregate network G' w.r.t. A'. We define possible aggregations upon multidimensional networks using Regular Expressions. In particular, $Q = \{q_1, q_2, ..., q_n\}$ is a set of n process cubes, where each q_i is a process cube, a placeholder for set of related entities and/or relationships among them, and can be encoded using regular expressions. In this context, each process cube q_i can extensively support multiple information needs with the graph data model (e.g., Definition 1) and one algorithm (regular language reachability). The set of related process cubes Q is designed to be customizable by local domain experts (who have the most accurate knowledge about their requirement) to codify their knowledge into regular expressions. These expressions can describe paths through the nodes and edges in the attributed graph: Q can be constructed once and can be reused for other processes. The key data structure behind the process cube is the Process Knowledge Graph, i.e., a graph of typed nodes, which represent process related entities (such as process instances, models, artifacts, actors, data sources, and information items), and typed edges, which label the relationships of the nodes to one another, illustrated in Fig. 2(A). We leveraged the graph mining algorithms in our previous work [5] to walk the graph from one set of interesting entities to another via the relationship edges and discover which entities are ultimately transitively connected to each other, and group them in folder nodes (set of related entities) and path nodes (set of related patterns). We use correlation-conditions [18] to partition the Process Knowledge Graph based on set of dimensions coming from the attributes of node entities. We use a path-condition [5] as a binary predicate defined on the attributes of a path that allows to identify whether two or more entities are related through that path.

Definition 5 (Dimensions)**.** Each process cube q_i has a set of dimensions $D = \{d_1, d_2, ..., d_n\}$, where each d_i is a dimension name. Each dimension d_i is represented by a set of elements (E) where elements are the nodes and edges of the Process Knowledge Graph. In particular, $E = \{e_1, e_2, ..., e_m\}$ is a set of m elements, where each e_i is an element name. Each element e_i is represented by a set of attributes (A), where $A = \{a_1, a_2, ..., a_p\}$ is a set of p attributes for element e_i, and each a_i is an attribute name. A dimension d_i can be

considered as a given query that require grouping graph entities in a certain way. Correlation-conditions and path-conditions can be used to define such queries.

A dimension uniquely identifies a subgraph in the Process Knowledge Graph, which we call a *Summary*. Now, we introduce the new notion of Narrative.

Definition 6 (Narrative)**.** A narrative $N = \{S, R\}$, is a set summaries $S = \{s_1, s_2, ..., s_n\}$ and a set of relationships $R = \{r_1, r_2, ..., r_m\}$ among them, where s_i is a summary name and r_j is a relationship of type 'part-of' between two summaries. This type of relationship enables the zoom-in and zoom-out operations (see Fig. 2(C)) to link different pieces of a story and enable the analyst to interact with narratives. Each summary $S = \{Dimension, View - Type, Provenance\}$, identified by a unique dimension D, relates to a view type VT (e.g., process, actor or data view) and assigned to a Provenance code snippet P to document the evolution of the summary over time (more nodes and relationships can be added to the Process Knowledge Graph over time). We leverage our work [6] to document the evolution of summaries over time.

The formalism of the summary S will enable to consider different dimensions and views of a narrative, including the event structure (narratives are about something happening), the purpose of a narrative (narratives about actors and artifacts), and the role of the listener (narratives are subjective and depend on the perspective of the process analyst). Also, it considers the importance of time and provenance as narratives may have different meanings over time. We develop a scalable summary generation algorithm and support three types of summaries. Figure 3 illustrates the scalable summary generation process. Following we introduce these summaries:

- Entity Summaries: We use correlation conditions to summarize the Process Knowledge Graph based on set of dimensions coming from the attributes of node entities. In particular, a correlation condition is a binary predicate defined on the attributes of attributed nodes in the graph that allows to identify whether two or more nodes are potentially related. Algorithm 1 in Fig. 3, will generate all possible entity summaries. For example, one possible summary may include all related images captured in the same location. Another summary may include all related images captured in the same timestamp.
- Relationship Summaries: We use correlation conditions to summarize the Process Knowledge Graph based on set of dimensions coming from the attributes of attributed edges. Algorithm 2 in Fig. 3, will generate all possible relationship summaries. For example, one possible summary may include all related relationships typed controlled-by and have the following attributes "Controlled-by (role = 'Investigator'; time = 'τ_1'; location = '255.255.255.0')". In the relationship summaries, we also store the nodes from and to the relationship, e.g., in this example the process instance and the actor.
- Path Summaries: We use path conditions to summarize the Process Knowledge Graph based on set of dimensions coming from the attributes of nodes

and edges in a path, where a path is a transitive relationship between two entities showing a sequence of edges from the start entity to the end. In particular, a path condition defined on the attributes of nodes and edges that allows to identify whether two or more entities (in a given Process Knowledge Graph) are potentially related through that path. Algorithm 3 in Fig. 3, will generate all possible path summaries. For example, one possible relationship summary includes all related images captured in the same location and contain the same information item, e.g., the missing person. Another relationship summary includes all related Tweets or emails sent on timestamp τ_1 and include the keyword Maisie (the missing person).

3.4 Process Analytics

In this phase, we present a spreadsheet like interface on top of the scalable summary generation framework. The goal is to enable analysts to interact with the narratives and control the resolutions of summaries. A narrative N can be analyzed using three operations: (i) roll-up: to aggregate summaries by moving up along one or more dimensions, and to provide a smaller summary with less details. (ii) drill-down: to disaggregate summaries by moving down dimensions; and to provide a larger summary with more details; (iii) slice-and-dice: to perform selection and projection on snapshots. To achieve this goal, we use the notion of spreadsheets and organize all the possible summaries in the rows and columns of a grid. Each tab in the spreadsheet defines a summary type (e.g., entity, relationship or path summary), the rows in a tab are mapped to the dimensions (e.g., Attributes of an entity), and the columns in a tab are mapped to various data islands in the Data Lake. Each cell will contain a specific summary.

We make a set of machine learning algorithms available as a service and to enable the analysts to manipulate and use the summaries in spreadsheets to support: (i) roll-up: the roll-up operation performs aggregation on a spreadsheet tab, either by climbing up a concept hierarchy (i.e., rows and columns which represent the dimensions and data islands accordingly) or by climbing down a concept hierarchy, i.e., dimension reduction; (ii) drill-down: the drill-down operation is the reverse of roll up. It navigates from less detailed summaries to more detailed summaries. It can be realized by either stepping down a concept hierarchy or introducing additional dimensions. For example, in Fig. 4, applying the drill-down operation on the cell intersecting time (dimension) and CCTV1 (data source) will provide a more detailed summary, grouping all the items over different points in time. As another example, applying the drill-down operation on the cell intersecting country (dimension) and Twitter (data source) will provide a more informative summary, grouping all the tweets, twitted in different counties; and (iii) slice-and-dice: the slice operation performs a selection on one dimension of the given tab, thus resulting in a sub-tab. The dice operation defines a sub-tab by performing a selection on two or more dimensions. This will enable analyst, for example to see Tweets coming from 2 dimensions such as time and

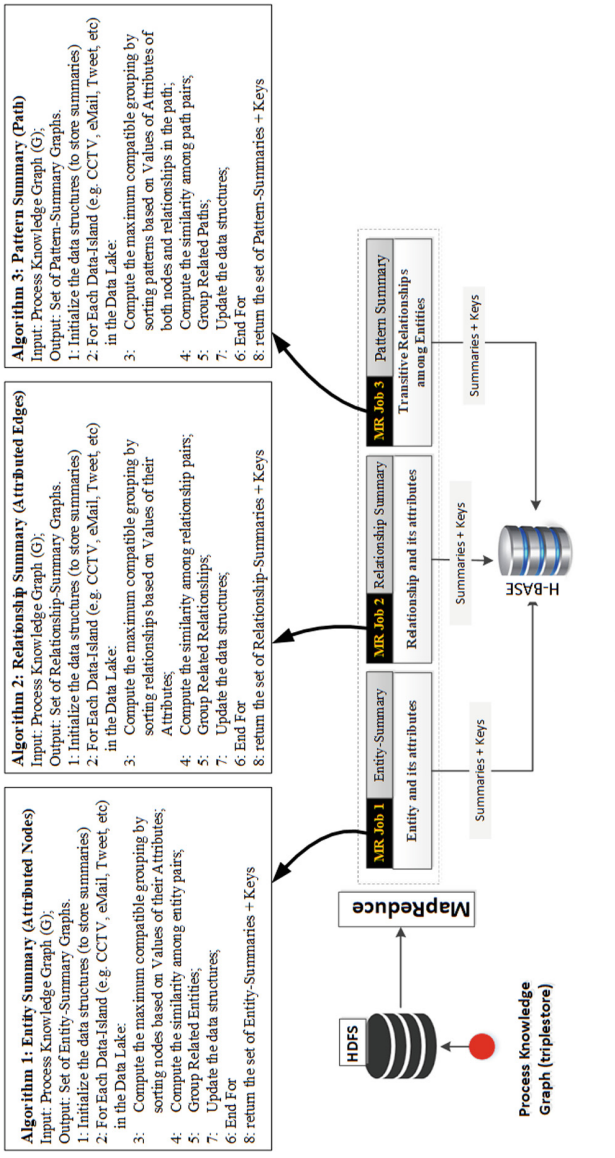

Fig. 3. Scalable summary generation.

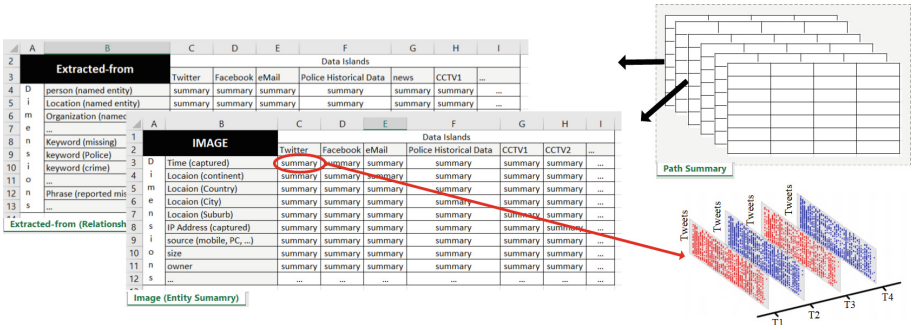

Fig. 4. Presenting a spreadsheet like interface on top of the scalable summary generation framework.

location. The slice-and-dice operation can be simply seen as a regular expression which groups together different entity and/or relationship summaries (presented in the spreadsheet tabs) and weaves them together to construct path summaries, illustrated in Fig. 4.

4 Implementation and Evaluation

We focus on the motivating scenario, to assist knowledge workers in the domain of law enforcement collect information from the investigation scene as well as the IoT-enabled devices of interest in an easy way and on a mobile device. The goal here is to contribute to research and thinking towards making the police officers more effective and efficient at the front-line, while augmenting their knowledge and decision management processes through Information and Communication Technology. We develop ingestion services to extract the raw data from IoT devices such as CCTVs, location sensors in police cars and smart watches (to detect the location of people on duty) and police drones. These services will persist the data in the data lake. Next and inspired by Google Knowledge Graph (developers.google.com/knowledge-graph/), we focused on constructing a policing process knowledge graph: an IoT infrastructure that can collaborate with internet-enabled devices to collect data, understand the events and facts and assist law enforcement agencies in analyzing and understanding the situation and choose the best next step in their processes. There are many systems that can be used at this level including our previous work (Curation APIs) [9], Google Cloud Platform (cloud.google.com/), and Microsoft Computer Vision API (azure.microsoft.com/) to extract information items from artifacts (such as emails, images, social items).

We have identified many useful machine learning algorithms and wrapped them as services to enable us to summarize the constructed knowledge graph, and to extract complex data structures such as timeseries, hierarchies, patterns and subgraphs and link them to entities such as business artifacts, actors, and activities. Figure 5, illustrates the taxonomy of these services. We use a

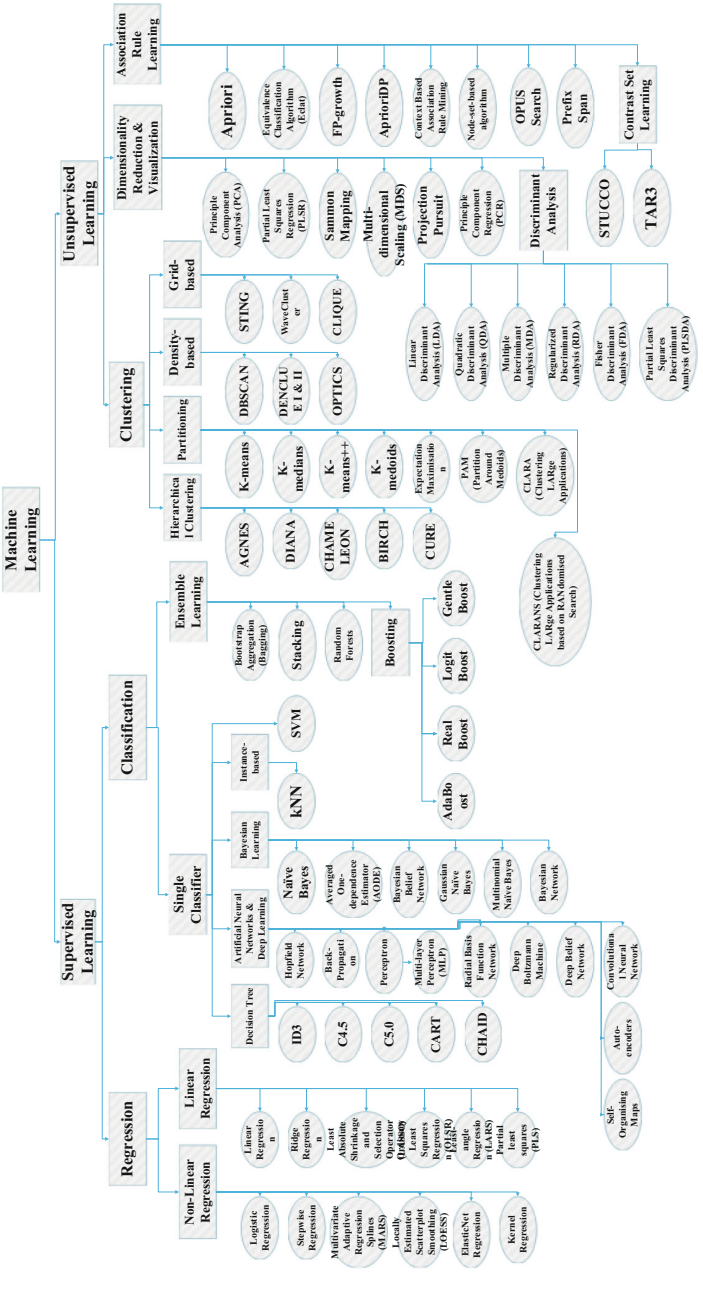

Fig. 5. The Taxonomy of the Machine Learning algorithms used as a Service to enable the knowledge workers interact with the summaries in an easy way.

spreadsheet-like dashboard that enables the knowledge workers interact with the summaries in an easy way. The dashboard enables monitoring the entities (e.g., IoT devices, people, and locations) and dig for the facts (e.g., suspects, evidences and events) in an easy way. A set of services has been developed to link the dashboard to the knowledge graph and the data summaries. A **demonstration** of the prototype can be found in: https://github.com/unsw-cse-soc/CoreKG.

The evaluation of accuracy and performance of the Data Lake and knowledge extraction services demonstrated in [3,9]. Figure 6 shows the performance of our access structure as a function of available memory for entity/relationship and path summaries. These summaries have been generated from a Tweet dataset having over 15 million tweets, persisted and indexed in the MongoDB (mongodb.com) database in our Data Lake. For the path summaries, we have limited the dept of the path to have maximum of three transitive relationship between the starting and ending nodes. The experiment were performed on Amazon EC2 platform using instances running Ubuntu Server 14.04. The memory size is expressed as a percentage of the size required to fit the largest partition of data in the hash access structure in physical memory. For efficient access to single cells (i.e. a summary) we built a partition level hash access structure where the partitions will be kept in memory and the operations will evaluated for one partition at a time. If a summary does not fit in memory we incur an I/O if a referenced cell is not cached. In the case of entity/relationship summary Fig. 6(A), this occurs when the available memory is around 40% of the largest summary, and for the path summary Fig. 6(B) this occurs when the available memory is around 30% of the largest summary.

As future work, we will evaluate the usability of the approach regarding the intended application audience, i.e., the police and expert users.

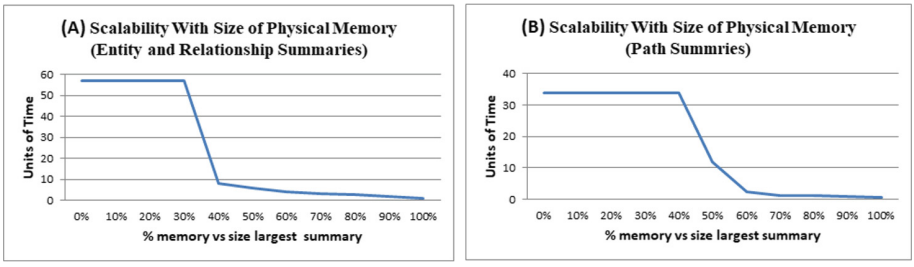

Fig. 6. Scalability with size of physical memory for entity and relationship summaries (A) and scalability with size of physical memory for path summaries (B).

5 Conclusion and Future Work

The large amount of raw data generated by IoT-enabled devices provide real-time intelligence to organizations which can enhance knowledge intensive processes.

For example, one of the interventions that have emerged as a potential solution to the challenges facing law enforcement officers is interactive constable on patrol system. In such a system, Internet-enabled devices and a mobile application that delivers policing capabilities to front-line officers (to make the work of the force more efficient and appropriate) plays an important role. Such an application improves knowledge exchange, communication practices, and analysis of information within the police force. To achieve this goal, in this paper, we present a scalable and extensible IoT-Enabled Process Data Analytics Pipeline (namely *iProcess*) to enable analysts to ingest data from IoT devices, extract knowledge from this data and link them to process (execution) data. To enhance the real-time dashboard, as a future work, we are working on a novel Platform-as-a-Service that makes it easy for developers of all skill levels to use machine learning technology, the way people use spreadsheet.

References

1. Bandyopadhyay, D., Sen, J.: Internet of Things: applications and challenges in technology and standardization. Wirel. Pers. Commun. **58**(1), 49–69 (2011)
2. Beheshti, A., Benatallah, B., Nezhad, H.: ProcessAtlas: a scalable and extensible platform for business process analytics. Softw. Pract. Exper. **48**(4), 842–866 (2018)
3. Beheshti, A., Benatallah, B., Nouri, R., Chhieng, V.M., Xiong, H., Zhao, X.: Coredb: a data lake service. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, 06–10 November 2017, pp. 2451–2454 (2017)
4. Beheshti, A., Benatallah, B., Nouri, R., Tabebordbar, A.: CoreKG: a knowledge lake service. In: Proceedings of the VLDB Endowment (PVLDB 2018), vol. 11(12) (2018). https://doi.org/10.14778/3229863.3236230
5. Beheshti, S., Benatallah, B., Motahari-Nezhad, H.R.: Scalable graph-based OLAP analytics over process execution data. Distrib. Parallel Databases **34**(3), 379–423 (2016)
6. Beheshti, S., Benatallah, B., Nezhad, H.R.M.: Enabling the analysis of cross-cutting aspects in ad-hoc processes. In: CAiSE, pp. 51–67 (2013)
7. Beheshti, S.-M.-R., Benatallah, B., Motahari-Nezhad, H.R., Sakr, S.: A query language for analyzing business processes execution. In: Rinderle-Ma, S., Toumani, F., Wolf, K. (eds.) BPM 2011. LNCS, vol. 6896, pp. 281–297. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23059-2_22
8. Beheshti, S., et al.: Process Analytics - Concepts and Techniques for Querying and Analyzing Process Data. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-25037-3
9. Beheshti, S., Tabebordbar, A., Benatallah, B., Nouri, R.: On automating basic data curation tasks. In: WWW (2017)
10. Benson, D.: The police and information technology. In: Technology in Working Order: Studies of Work, Interaction, and Technology, pp. 81–97 (1993)
11. Bhattacharya, K., Gerede, C.E., Hull, R., Liu, R., Su, J.: Towards formal analysis of artifact-centric business process models. In: BPM, pp. 288–304 (2007)
12. Braga, A.A., Weisburd, D.L.: Police innovation and crime prevention: lessons learned from police research over the past 20 years (2015)

13. Carey, M.J., Onose, N., Petropoulos, M.: Data services. Commun. ACM **55**(6), 86–97 (2012)
14. Casati, F., Castellanos, M., Dayal, U., Salazar, N.: A generic solution for warehousing business process data. In: Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 1128–1137. VLDB Endowment (2007)
15. Da Xu, L., He, W., Li, S.: Internet of Things in industries: a survey. IEEE Trans. Industr. Inf. **10**(4), 2233–2243 (2014)
16. Gerede, C., Su, J.: Specification and verification of artifact behaviors in business process models. In: ICSOC, pp. 181–192 (2007)
17. Kuo, J.: A document-driven agent-based approach for business processes management. Inf. Softw. Technol. **46**(6), 373–382 (2004)
18. Motahari-Nezhad, H.R., Saint-Paul, R., Casati, F., Benatallah, B.: Event correlation for process discovery from web service interaction logs. VLDB J. Int. J. Very Large Data Bases **20**(3), 417–444 (2011)
19. Ngu, A.H.H., Gutierrez, M.A., Metsis, V., Nepal, S., Sheng, Q.Z.: IoT middleware: a survey on issues and enabling technologies. IEEE Internet Things J. **4**(1), 1–20 (2017)
20. Reijers, H., Rigter, J., Aalst, W.: The case handling case. Int. J. Cooperative Inf. Syst. **12**(3), 365–391 (2003)
21. Schonenberg, H., Weber, B., van Dongen, B.F., van der Aalst, W.M.P.: Supporting flexible processes through recommendations based on history. In: BPM, pp. 51–66 (2008)
22. Sun, Y., Song, H., Jara, A.J., Bie, R.: Internet of Things and big data analytics for smart and connected communities. IEEE Access **4**, 766–773 (2016)
23. Sun, Y., Su, J., Yang, J.: Universal artifacts: a new approach to Business Process Management (BPM) systems. ACM Trans. Manage. Inf. Syst. **7**(1), 3:1–3:26 (2016)
24. van der Aalst, W., et al.: Process mining manifesto. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) BPM 2011. LNBIP, vol. 99, pp. 169–194. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28108-2_19