Soundness and Ontology-Based Consistency of Sensor Data Acquisition Plans

Luca Ferrari(✉), Marco Mesiti(✉), and Stefano Valtolina(✉)

Department of Computer Science "Giovanni Degli Antoni",
University of Milano, Milan, Italy
{lferrari,mesiti,valtolina}@di.unimi.it

Abstract. The verification of the soundness and consistency of data acquisition plans is an important requirement in the loading of data generated by physical and social devices. In this paper we discuss these properties in the context of the StreamLoader system.

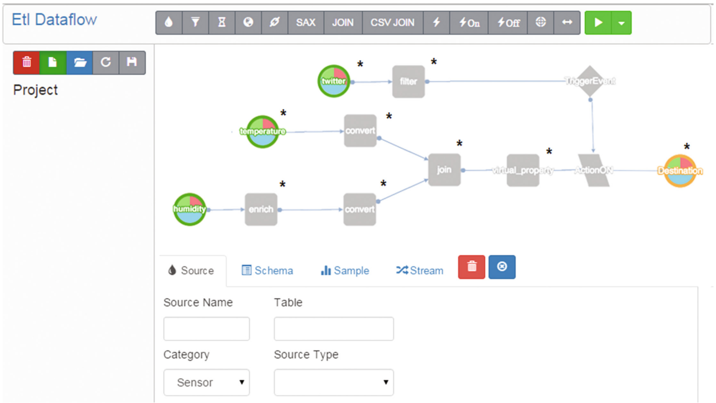
Keywords: Internet of Things · Data streaming techniques · Distributed data integration and semantic enrichment · IoT platforms for big data

1 Introduction

Current workflow management systems (WfMS) provide users with various facilities for simplifying the development of complex programs by means of graphical interfaces. Commercial systems such as Talend Studio (www.talend.com), StreamBase Studio (www.streambase.com), Waylay.io (www.waylay.io) are designed for offering programming assistance in the design of workflows/dataflows as graphs of connected nodes representing tasks and data sources. These environments support the full application lifecycle, spanning feed integration, application modeling, development, data streams recording, testing, and debugging.

According to this design strategy, we are working on a web-based system, named StreamLoader [5], that offers facilities for the development of data acquisition plans specifically tailored for heterogeneous sensor data through the definition of a graph of services that load, filter, transform, aggregate, and compose different kinds of stored and stream data. Unlike the WfMS, StreamLoader is designed for guaranteeing the specification of dataflows that can be soundly execute, and that the generated data are semantically consistent w.r.t. a *Domain Ontology* (*DO*). A *DO* is used for describing the capabilities and properties of sensors, the act of sensing and the resulting observations in a specific domain (e.g. the analysis of meteorological conditions). In this type of analysis, the ontology should represent sensor types, location, acquisition time, and thematic definitions (e.g. the type of acquired data and its accuracy, precision, and measurement range). To do it, the SSN Ontology [3,4] has been extended with concepts/relationships that are usually adopted in the specific domain and the obtained DO can be used as constraints in the formulation of data acquisition

P. Ciancarini et al. (Eds.): EKAW 2016 Satellite Events, LNAI 10180, pp. 109–113, 2017.
DOI: 10.1007/978-3-319-58694-6_11



Source Name

Table

Category

Sensor

Source Type

Fig. 1. Main screen of the StreamLoader web application

plans. StreamLoader adopts a very flexible, multi-granular, spatio-temporal-thematic (*STT*) data model that allows to consider heterogeneous streams of events (and stored data) generated by different kinds of sensors.

In this paper we focus on describing the set of services that can be applied for defining the dataflow and how their combination can be considered sound (w.r.t. the STT data model) and consistent (w.r.t. the adopted DO). These concepts are fundamental to guarantee that dataflows can be computed without errors, and the produced data complies with the DO along its STT dimensions.

2 Data Plan Specification

The management of complex events has been widely discussed in the past [1]. From our point of view, an event is a record of an observed change of state in the monitored situation at a given point in time. Each atomic event is characterized by the information about when it happened (time dimension), where it happened (space dimension), and what it concerned (thematic dimension). Starting from these atomic events, complex events can be generated that point out complex correlations among the basic events. To compose basic events, a set of services can be applied. Services are based on operations for the application of filters, transformation, aggregation, and composition. Operations can be classified in non-blocking (filter, cull-time/space, enrich, virtual property, transform) and blocking (aggregation, union, join, trigger on/off, convert). The formers are directly applied on each tuple when they are processed, whereas the others are processed according to time-based windows. In applying these services, we consider spatial and temporal types at different granularities that can be exploited for the specification of sound plans (further details in [5]).

Through the visual interface in Fig. 1, StreamLoader allows users to drag and drop icons representing sensors on a canvas and connect them by using the operations made available for the specification of the dataflow. In the depicted situation, when the average number of tweets in the last hour about hot temperature is greater than 20, the apparent temperature is calculated by considering temperatures and humidities identified in the city zones. Since the corresponding devises gather events at different temporal granularities, the need arises to convert them at hour granularity. Moreover, the enrich operation allows to add spatio-temporal information of where the humidity information is acquired.

3 Sound and Consistent Specification

Each service used in a dataflow is characterized by several constraints that are exploited for the verification of the service applicability that take into account the use of the spatio-temporal granularities. According to this consideration a data plan is *sound* if for each applied service the number of input streams is equal to the number of expected input streams; the parameters are specified; and, the applicability conditions on the input stream are verified.

The DO is generated by the domain experts by specifying a precise meaning of the spatio-temporal-thematic dimensions that are used in their context. This means that the structural part of the Ontology contains the spatial/temporal granularities that are adopted, the thematics that are recognized along with their attributes and types. Moreover, instances are included for representing what things should be looked up to support the creation of meaningful streams that is, sensors, observations, and related concepts.

A sound data plan is *consistent* when the schema of the output stream of generated events is consistent w.r.t. the DO. Therefore, in our environment is tolerated that the input streams or internal operations generate streams that are not consistent w.r.t. the DO, but the final stream should be consistent. This definition allows to face the issue that some devices may not produce events according to the STT dimensions, but the operations contained in the dataflow can transform the stream in one whose semantics is well-described in the DO. For examples the temperature sensors that are disseminated in a zone of the city can produce simply the observed values with no information about the time and position of the sensors. However, the gateway in charge of acquiring their observations can calculate the average temperature once a hour and assigns its position as location (at the `zone` spatial granularity) and the current hour as temporal dimension (at the `hour` temporal granularity).

Each time a sensor is virtualized in our environment, its data schema is mapped to the concepts of the DO along with the STT dimensions. If all dimensions are specified and properly mapped on the Ontology concepts, we can consider the produced event stream consistent. However, we also accept data sources for which the consistency is not verified. Indeed, other operations can be applied on them to make them consistent. Whenever the user needs to create a data acquisition plan, she introduces in the canvas icons representing the sensors and the services for their manipulation. Whenever the graph corresponding to the current dataflow is sound according to our definition, the consistency is checked starting from the data sources and moving toward the final node that collects the output data stream. For each node of the graph, new instances can be introduced in the DO (when required) in order to maintain their description at the Ontological level. Specifically, the services filter, cull-time/space do not modify the data model and then the Ontology Instances are left unchanged. By contrast, for the other services (transform, enrich, virtualProperty, aggregation, union, join, convert and trigger) it is necessary to apply a set of instructions for modifying the Ontology Instances according to the service specification. For their proper handling, a virtual sensor is introduced whose properties depend on the applied operator and on the input streams. The virtual sensor is obtained in two steps. First, the incoming sensor is cloned and renamed. Then, the cloned instance and its links are modified to comply with the operator specification.

If the initial dataflow is consistent, since each service adds or clones existing instances or associations, these services do not jeopardize the consistency with respect to the DO. This is guaranteed by the fact that, if one of the services of enrich, virtualProperty and transform creates an instance of a class not foreseen in the DO, the service fails to modify the Ontology. In the other cases, the application of the services aggregation, join, trigger, union, convert does not affect the consistency of the Ontology because they only add new associations or instances of existing classes. However, if the initial consistency of the sources is not verified, it can become consistent after the application of other operations. In Fig. 1, a * denotes that a service is consistent w.r.t. the DO. The dataflow is consistent even if some of the nodes of the graph are not consistent.

4 Conclusions

In this paper we discussed the concepts of sound and consistent specification of data acquisition plans for streams of sensor data. Once the dataflow is specified it is executed on a distributed system during the transfer of the data from the sources to the destination [2]. In the poster we will provide an overview of the StreamLoader architecture and the adopted data model. Then, we will present the adopted services with their meaning and constraints. Finally we will discuss examples of data acquisition plans that are only sound or sound and consistent with respect to a given DO in the context of meteorology. Future works aim at studying ontology-based approaches supporting stream reasoning. The idea is to help domain experts in detecting possible operators to apply in StreamLoader for enriching inconsistent sensor data specifications. In this way we can use our *DO* for providing explicit semantics for events but at the same time, for reasoning on the correctness of semantic integrations and extensions.

References

1. Cugola, G., Margara, A.: Processing flows of information: from data stream to complex event processing. ACM Comput. Surv. **44**(3), 62 pages (2012). Article No. 15
2. Zettsu, K.: Service-controlled networking. J. Nat. Inst. Inf. Commun. Technol. **62**(2), 177–184 (2015)
3. W3C Semantic Sensor Network Group: Semantic Sensor Network Ontology (2005)
4. Compton, M., et al.: The SSN ontology of the W3C semantic sensor network incubator group. Web Semant. Sci. Serv. Agents World Wide Web **17**, 25–32 (2012)
5. Mesiti, M., et al.: StreamLoader: an event-driven ETL system for the on-line processing of heterogeneous sensor data. In: Proceedings of International Conference on Extending Database Technology, pp. 628–631 (2016)