Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Knowledge-infused and consistent Complex Event Processing over real-time and persistent streams

Qunzhi Zhou[a,*], Yogesh Simmhan[b], Viktor Prasanna[c]

[a] Department of Computer Science, University of Southern California, Los Angeles, USA
[b] Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India
[c] Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, USA

HIGHLIGHTS

- A semantic CEP model is introduced to query across real-time and persistent streams.
- The model's analytic capability is illustrated using case studies from Smart Grid.
- Approaches to translate the model into scalable execution are discussed and evaluated.

ARTICLE INFO

Article history:
Received 1 December 2015
Received in revised form
23 June 2016
Accepted 27 October 2016
Available online xxxx

Keywords:
Complex Event Processing
Query optimization
Consistency
Internet of Things (IoT)
Big data
Semantic web
Knowledge-base
Real-time processing
Reliability

ABSTRACT

Emerging applications in Internet of Things (IoT) and Cyber–Physical Systems (CPS) present novel challenges to Big Data platforms for performing online analytics. Ubiquitous sensors from IoT deployments are able to generate data streams at high *velocity*, that include information from a *variety* of domains, and accumulate to large *volumes* on disk. Complex Event Processing (CEP) is recognized as an important real-time computing paradigm for analyzing continuous data streams. However, existing work on CEP is largely limited to relational query processing, exposing two distinctive gaps for query specification and execution: (1) infusing the relational query model with higher level *knowledge semantics*, and (2) seamless *query evaluation across temporal spaces* that span past, present and future events. These allow accessible analytics over data streams having properties from different disciplines, and help span the velocity (real-time) and volume (persistent) dimensions. In this article, we introduce a Knowledge-infused CEP (χ-CEP) framework that provides *domain-aware knowledge query constructs* along with temporal operators that allow end-to-end queries to span across *real-time and persistent streams*. We translate this query model to efficient query execution over online and offline data streams, proposing several optimizations to mitigate the overheads introduced by evaluating semantic predicates and in accessing high-volume historic data streams. In particular, we also address temporal consistency issues that arise during fault recovery of query plans that span the boundary between real-time and persistent streams. The proposed χ-CEP query model and execution approaches are implemented in our prototype semantic CEP engine, *SCEPter*. We validate our query model using domain-aware CEP queries from a real-world Smart Power Grid application, and experimentally analyze the benefits of our optimizations for executing these queries, using event streams from a *campus-microgrid IoT deployment*. Our results show that we are able to sustain a processing throughput of 3, 000 events/secs for χ-CEP queries, a 30× improvement over the baseline and sufficient to support a Smart Township, and can resume consistent processing within 20 secs after stream outages as long as 2 hours.

1. Introduction

There is a growing prevalence of Internet of Things (IoT) sensors and actuators, both for specific domains such as Smart Grids and Smart Transportation, and through lifestyle devices such as Smart Watches and Fitness Bands. These sensors generate streams

* Corresponding author.
E-mail address: zhouqunzhi@gmail.com (Q. Zhou).

http://dx.doi.org/10.1016/j.future.2016.10.030
0167-739X/

of events that arrive continuously, and can include observations from multiple domains that need to be analyzed. Complex Event Processing (CEP) is a computing paradigm for online analytics over such high velocity data streams [1]. Contemporary CEP systems offer the capability to specify event patterns to detect value thresholds or correlation constraints, and execute them continuously over events streams. In particular, CEP addresses the *velocity* dimension of the 3-Vs of Big Data [2], *volume* and *variety* being the other two, and has grown popular for operational intelligence where online pattern detection drives real-time response. It has been used in domains varying from mobile computing [3] and financial services [4] to healthcare [5] and sports analytics [6].

One emerging domain where CEP can prove vital is in *Cyber–Physical Systems* (CPS) [7], which is a special case of IoT. In CPS, the operation and optimization of *physical* infrastructure is based on analytics performed on *cyber*-infrastructure, and typically happens in a closed-loop cycle. CPS encompasses many aspects of smart cities where diverse events on infrastructure conditions, be they about a transportation network [8] or power grid [9], are integrated with prior knowledge of the infrastructure to offer insight on the system behavior. CEP engines can help analyze such event streams and detect event patterns that need an operational response. For e.g., detecting a traffic overflow based on events from sensors on an upstream road can cause downstream traffic signals to change, or stress on a neighborhood transformer detected by analyzing residential smart meter readings can trigger a notification requesting consumers in that community to curtail their power consumption.

While CEP offers a useful paradigm to perform real-time analytics over event streams, there are two distinctive capabilities lacking in traditional CEP engines that are necessary for their effective use by emerging IoT applications:

1. IoT domains need to perform analytics over multi-disciplinary data sources for effective decision making. A CEP query model has to be expressive enough to capture such information richness while also being simple enough for end users to specify such analytics by hiding domain complexities. Traditional CEP systems require *syntactic* queries to be specified over explicit properties present in the event contents, such as sensor IDs and equipment numbers. This makes CEP queries difficult to specify and manage for many numbers of diverse devices and sensors that are constantly in flux. *Semantic* concepts defined by domain ontologies [10,11] can help raise the abstraction by referring to events using concepts rather than just content. This requires the CEP query definitions to be infused with knowledge models and *semantic predicates*, and further translated into *efficient execution*.
2. While CEP systems allow queries to be specified over current and future events that arrive on an event stream, event analytics may require the correlation between events that happen in the present as well as in the past. Traditional CEP systems do not allow queries to be specified *after* an event has occurred to match the past event, necessary for exploratory analytics. Secondly, even after a query is specified, CEP systems can be memory-intensive when processing queries with a large time window, some as wide as days. This motivates the need for *lazy definition* of queries after an event has happened, and their *consistent and scalable execution* over end-to-end event streams that span past events persisted to disk and real-time events that arrive over the network.

In other words, while current CEP systems support analytics over high *velocity* event data, we also need to support data *variety* in the form of diverse domains concepts present in IoT event streams, and analysis over large *volumes* of archived event streams, thus uniquely encompassing all three dimensions of Big Data.

Most existing CEP systems expose gaps in their ability to model queries at a higher-abstraction, and their execution on end-to-end event streams. These systems process relational events with syntactic queries directly defined on their properties, and expose users to the underlying events' structural heterogeneity [1]. Recently, C-SPARQL [12] and ETALIS [13] introduced event context or semantics into CEP for abstract query specification, allowing background knowledge to be combined with real-time events. But these are solutions that rely on Semantic Web technologies, leveraging inference engines to model and query over events and domain knowledge. As a result, common CEP temporal patterns such as klenee closure and matching policies for event selection and consumption [14,15] are not supported. Further they lack the scalability of CEP systems that are optimized for pattern matching over streaming events. On the other hand, existing CEP systems focus on processing real-time events, without considering archived event streams. Integrated querying over real-time and persistent data has attracted some interest from active databases [16]. These leverage triggers in relational query engines to process time varying data that is persisted. DataCell [17] layers in-memory tables on top of database kernels to handle online queries. These *database-centric* systems sacrifice the expressivity of CEP temporal query patterns and introduce additional latency into matched results. A recency-based CEP model [18] supports a *happened-before* relation which links live streams with persisted events. However, this is limited to correlating patterns between real-time and archived streams, rather than seamlessly detecting patterns that span across both. Big Data platforms that perform in-memory computing, like Apache Spark Streaming, support high-throughput incremental processing over persisted data by loading batches of datasets into memory. But Spark treats the data content as opaque and users need to implement all processing logic over them. Further, while such systems are faster than batch processing platforms like Hadoop, they still have a higher latency than CEP engines.

In this article, we propose *SCEPter*, a Knowledge-infused CEP (χ-CEP , pronounced *kai*-CEP) framework which uniformly processes queries across persistent and real-time event streams, end-to-end, and addresses the gaps identified above. *SCEPter* is motivated by and evaluated within the Smart Power Grid CPS domain, as part of the US Department of Energy sponsored Los Angeles Smart Grid project [19]. *SCEPter* allows users to specify expressive event patterns using semantic concepts over heterogeneous information sources, permits lazy-specification of queries for online and *post facto* analytics, ensures temporally consistent execution across end-to-end event streams, and includes optimizations to mitigate performance overheads introduced by such features.

Specifically, our contributions are as follows:

1. We take a CEP-centric approach to infuse knowledge-models and domain semantics into relational CEP events (Section 3). Further, we propose a unified χ-CEP query model that supports *semantic predicates* that leverage both real-time event data and static knowledge-bases, and *temporal predicates* that can operate end-to-end over real-time and persistent event streams.
2. We discuss query processing techniques for real-time event streams (Section 4), persistent event archives (Section 5), and executions that span the temporal boundary between the two (Section 6). We propose *performance optimizations* like event buffering and semantic caching for real-time querying, and hybrid rewriting and replay for archive processing to offer low latency, consistency and resiliency in the presence of temporal gaps in event streams.