Chapter 10
Stream and Event Processing Services for Real-time Linked Dataspaces

Keywords Stream processing · Event processing · Real-time query · Entity-centric · Lambda architecture

10.1 Introduction

The goal of Real-time Linked Dataspaces is to support a real-time response from intelligent systems to situations of interest within a smart environment by providing data processing support services that follow the data management philosophy of dataspaces and meet the requirements of real-time data processing. This part of the book details support services to process streaming and event data which support loose semantic integration and administrative proximity. Support services include entity-centric queries, complex event processing, stream dissemination, and semantic matching for heterogeneous events. The goal of these services is to support a real-time linked dataspace to get setup and running with a low overhead for administrative and semantic integration costs (e.g. establishing data agreements, service selection, and service composition).

Section 10.2 of this chapter lays out pay-as-you-go services in the context of event and stream processing. Section 10.3 details the entity-centric real-time query service, including its architecture, service-levels, and service performance, and the chapter concludes with a summary in Sect. 10.4.

10.2 Pay-As-You-Go Services for Event and Stream Processing in Real-time Linked Dataspaces

To support the interconnection of intelligent systems in the data ecosystem that surrounds a smart environment, there is a need to enable the sharing of data among systems. A data platform can provide a clear framework to support the sharing of data among a group of intelligent systems within a smart environment [1] (see

E. Curry, *Real-time Linked Dataspaces*,
https://doi.org/10.1007/978-3-030-29665-0_10

Chap. 2). In this book, we advocate the use of the dataspace paradigm within the design of data platforms to enable data ecosystems for intelligent systems.

Dynamic data from sensors and IoT devices comprise a significant portion of data generated in a smart environment. Responding to this trend requires a data platform to provide specific support services designed to work with real-time data sources. These services must keep with the dataspace philosophy; thus, they must co-exist, and co-evolve over time, and ensure a rigid data management approach does not subsume the source systems. Within the dataspace paradigm, data management pushes the boundaries of traditional databases in two main dimensions [2]: (1) Administrative Proximity: which describes how data sources within a space of interest are close or far in terms of control, and (2) Semantic Integration: which refers to the degree to which the data schemas within the data management system are matched up.

We have created the Real-time Linked Dataspace (RLD) (see Chap. 4) as a data platform for intelligent systems within smart environments. The RLD combines the pay-as-you-go paradigm of dataspaces with linked data and real-time stream and event processing capabilities to support a large-scale distributed heterogeneous collection of streams, events, and data sources [4]. The RLD follows a set of principles to describe the specific requirements within a real-time setting:

- An RLD must deal with many different formats of streams and events.
- An RLD does not subsume the stream and event processing engines; they still provide individual access via their native interfaces.
- Queries in the RLD are provided on a best-effort and approximate basis.
- The RLD must provide pathways to improve the integration between the data sources, including streams and events, in a pay-as-you-go fashion.

In order to enable these principles to support real-time data processing for events and streams, we explore new techniques to support approximate and best-effort stream and event processing services within an RLD-Support Platform (RLD-SP). The RLD-SP services support many formats of data, do not depend on prior-agreement for composition or dissemination, and provide a best-effort quality of service and approximate answers using a pay-as-you-go approach. As shown in Fig. 10.1, the stream and event processing services provided by the RLD-SP are:

- **Entity-Centric Index:** The entity-centric index enables unified queries across live streams, historical streams, and entity data to enable full entity-centric views of the current and past state of the smart environment.
- **Stream Dissemination:** A key challenge for an RLD-SP is to disseminate events and streams to relevant data consumers efficiently. The dataspace must facilitate machine-to-machine communications to build an efficient stream dissemination system for a smart environment.
- **Complex Event Processing:** The individual and compositions of event services within a smart environment can have different quality-of-service levels (e.g. latency, accuracy, reliability). An RLD-SP must support quality-of-service aware complex event service compositions to maximise the level of service available.

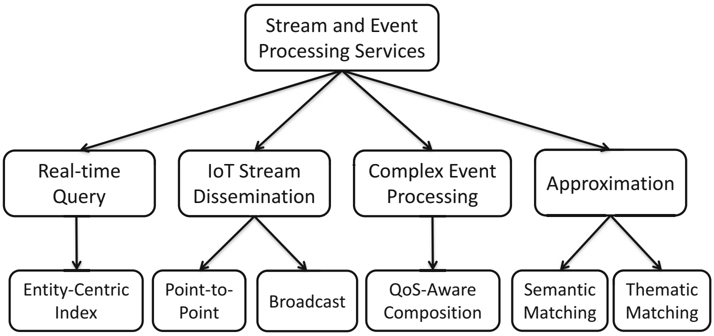

Fig. 10.1 Support for stream and event processing in the Real-time Linked Dataspace

Table 10.1 Pay-as-you-go stream and event support services in the RLD-SP [4]

Pay-as-you-go star rating	Entity-centric index	Complex event processing	Stream dissemination	Semantic approximation
* Basic	None	None	None	None
** Machine-readable	Basic processing	Single stream	None	Semantic matching
*** Basic integration	Historical views of streams	Multi-service composition	Point-to-point	Thematic matching
**** Advanced integration	Stream enrichment with context and entity data	Quality of service aware service composition	Wireless broadcast	Entity-centric matching
***** Full semantic integration, search, and query	Entity-centric real-time query	Context-aware composition	Complex patterns	Context-aware matching

- **Approximation:** RLD-SP needs to be able to support the processing of heterogeneous events. Semantic event matchers are one approach to handle data heterogeneity within real-time events when few or no prior-agreements exist.

Each of these support services is designed following a tiered model for service provision. This means that it can provide incremental support for participants in the RLD in a "pay-as-you-go" fashion. Table 10.1 identifies possible service-tiers available from each service aligned to the RLD 5 Star pay-as-you-go model. It should be noted that not all service-tiers have been implemented within each service. Rather, the implementation of the support services also follows an incremental approach with service-tiers developed on an as-needed basis based on the actual requirements of the different smart environments. In this light, Table 10.1 also serves as a roadmap for the development of each service capturing future work for some services.

In the remainder of this part of the book, we detail the above support services and focus on their role in supporting real-time data processing in dataspaces. Each chapter details the instantiation of a support service and its evaluation. The remainder of this chapter describes the entity-centric real-time query service.

10.3 Entity-Centric Real-time Query Service

An essential requirement in a smart environment is the querying of real-time data streams. Within the RLD [4], this is achieved by the entity-centric real-time query service that enables unified queries across live streams, historical streams, and entity data to enable full entity-centric views of the current and past state of the smart environment. This section first discusses the Lambda Architecture and then details how it has been extended to support entity-centric real-time queries.

10.3.1 Lambda Architecture

The Lambda Architecture is a frequently used Big Data processing architecture, which realises that both real-time and historical data analytics are crucial to support data analysis within smart environments. Rather than using two different systems for processing real-time data and historical data, the Lambda Architecture [268] allows seamless ingestion and processing of live and historical streaming data within a single architecture, as illustrated in Fig. 10.2.

Streams of events can be sourced from a variety of systems such as sensors, database logs, and website logs. All data entering the system is processed by both the batch layer and the speed layer. The batch layer pre-computes batch views of the stored raw data. The serving layer indexes the batch views for low-latency fast-access queries by applications. The speed layer deals with high-velocity updates by providing real-time append-only views of recent data. Queries are answered by merging results from both batch views (data-at-rest) and real-time views (data-in-motion). The Lambda Architecture has proved very useful for data management within smart environments [33].

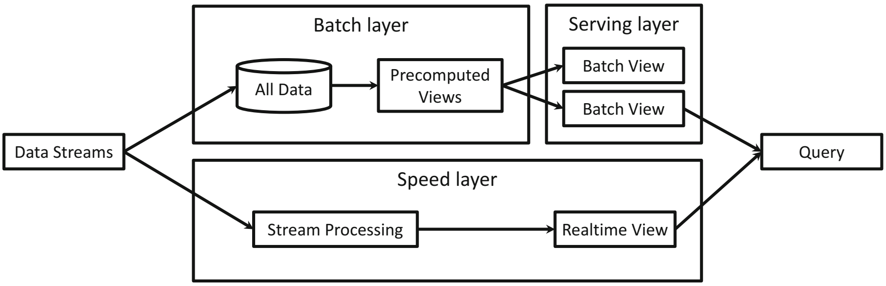

Fig. 10.2 The three layers of the Lambda Architecture

10.3.2 Entity-Centric Real-time Query Service

In terms of real-time data processing, the Lambda approach meets many of the data management requirements for intelligent system data ecosystems defined in Chap. 2. However, Lambda does not natively support the inclusion of entity and contextual data within the indexing and querying process. This means that applications need to maintain the relationship between the Lambda index and the entities in the dataspace by themselves. Ideally, an entity-centric real-time query service would be provided by the RLD-SP to remove the need for applications to manage the entity/stream relationship.

To meet this requirement, we designed an extension of the Lambda Architecture that includes the addition of an entity layer for the indexing of entity data alongside historical and live streams. The approach enables the serving layer to provide merged views across all three layers, removing the need for applications to maintain the entity/stream relationships. The entity-centric real-time query service is part of the RLD-SP and is tightly integrated with other support services such as the catalog, entity management, and access control services.

Figure 10.3 illustrates the design of the entity-centric real-time query service. The main components are:

- **Entity Data (Catalog):** Provides entity data from the catalog and entity management services.
- **Data Streams:** Produced by the "things" and sensors within the smart environment.
- **Batch Layer:** Provides batch-based processing for accurate, but delayed views of historical data.
- **Speed Layer:** Provides real-time processing for data with low-latency processing requirements. Streams in the speed layer are not stored but processed on-the-fly to guarantee low-latency approximate views of the data to complement the older

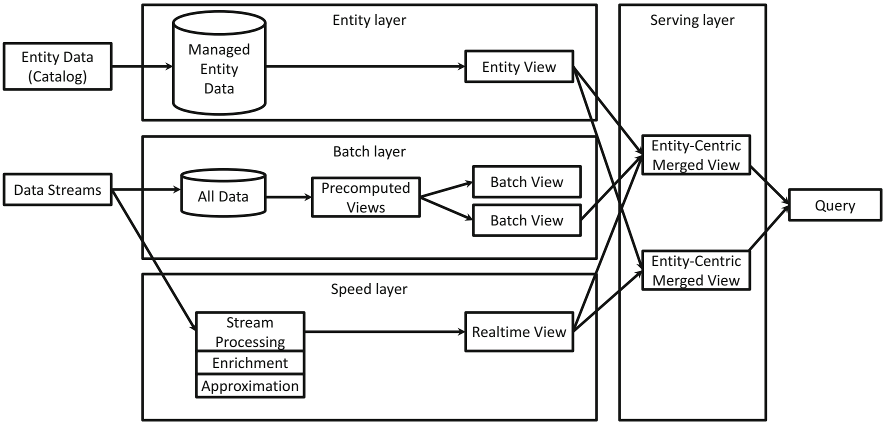

Fig. 10.3 The four layers of the entity-centric real-time query service [4]

views achieved by the batch layer. The speed layer provides several support services for processing event data, such as approximate event matching and event enrichment.

- **Entity Layer:** Provides a view of the managed entities within the RLD working closely with the catalog and entity management service.
- **Serving Layer:** Provides applications and users with a single entity-centric query interface for data access and query. This layer transparently splits queries to the batch, speed, and entity layers, and combines pre-computed views over the three layers.
- **Query:** Request for entity-centric views from applications, analytics, and users.

10.3.3 Pay-As-You-Go Service Levels

Dataspace support services follow a tiered approach to data management that reduces the initial cost and barriers to joining the dataspace. When tighter integration into the dataspace is required, it can be achieved incrementally by following the service tiers defined. The incremental nature of the support services is a core enabler of the pay-as-you-go paradigm in dataspaces. The tiers of service provision provided by the entity-centric real-time query service in the RLD follows the 5 Star pay-as-you-go model (detailed in Chap. 4). The service has the following tiered-levels of support:

1 Star **No Service:** Streams are not managed in the service.

2 Stars **Basic Processing:** Basic real-time stream processing in the speed layer only.

3 Stars **Historical Views:** Streams are stored in the batch layer for historical views.

4 Stars **Enrichment:** Streams are enriched with context and entity data from the catalog and entity management service.

5 Stars **Entity-Centric:** Streams are processed in all three layers to provide entity-centric real-time queries.

The unified entity-centric views provided by the service proved to be beneficial to developers/data scientists using the RLD; a performance assessment of the service is provided in the next section.

10.3.4 Service Performance

A key contribution of a real-time linked dataspace support platform is the entity-centric real-time query service for the RLD. The query latency for the service was evaluated within each environment to ensure it could support interactive user querying [269, 270]. We evaluated seven common queries within the developed applications to determine the level of query interactivity of the service. Table 10.2

Table 10.2 Query latency (seconds) of entity-centric real-time query service [4]

Query type	Airport A	Mixed use A	Home A	School A	Airport B	Mixed use B
timeBoundary	0.266415	0.076204	0.078805	0.076004	0.080605	0.078605
dataSourceMetadata	0.091405	0.078005	0.080805	0.153609	0.137808	0.092205
segmentMetadata	0.073804	0.084405	0.074004	0.140008	0.077405	0.077604
Search	0.162609	0.142808	0.085805	0.076404	0.136808	0.204812
Timeseries	0.072404	0.080005	0.077204	0.072404	0.134008	0.083605
groupBy	0.073404	0.078205	0.075604	0.081605	0.078605	0.072204
topN	0.078405	0.086805	0.072604	0.073004	0.077804	0.076604

presents these results based on the average of five runs of each query. Most queries have an “instantaneous response” of under 0.1 seconds, and all queries are responsive under 1 second, which is needed for “good navigation” [270]. This initial evaluation demonstrates the suitability of the query service for serving intelligent applications within smart environments.

10.4 Summary

The chapter provides a high-level overview of the real-time data processing support services within Real-time Linked Dataspaces (RLDs). Each service follows the pay-as-you-go data management philosophy of dataspaces. The goal of the services is to support an RLD to get setup and running with a low overhead for administrative and semantic integration costs (e.g. establishing data agreements, service selection, and service composition). This requires us to explore new techniques to support approximate and best-effort stream and event processing services which support loose semantic integration and administrative proximity. Specialised support services include complex event processing, event service composition, stream dissemination, stream matching, and approximate semantic matching. The chapter details the entity-centric real-time query service, including its architecture and service performance.