Review of “A Real-time Linked Dataspace for the Internet of Things: Enabling ‘‘Pay-As-You-Go’’ Data Management in Smart Environments”

Abdur Rafay Saleem
University of Central Florida
ab464825@ucf.edu

***Abstract*—This is a review of the paper “A Real-time Linked Dataspace for the Internet of Things: Enabling ‘‘Pay-As-You-Go’’ Data Management in Smart Environments” by Edward Curry and others. It is divided into three sections: a description of the research problems and the paper's technical contributions, a detailed breakdown of the proposed techniques, and a critical assessment of their effectiveness.**

I. RESEARCH PROBLEM

Smart environments, a popular phenomenon in today's era of technological advancement, require effective data management to handle the influx of real-time and large-scale data. However, there are challenges highlighted by the paper. Integrating new digital systems with traditional information infrastructures is a key challenge. Additionally, managing information availability and access among stakeholders is another hurdle. It is also important to analyze real-time data in conjunction with the problem context or industrial practices to derive meaningful insights. Finally, scalability and cost-effectiveness are crucial. The paper proposes considering a Dataspace as a potential solution to these challenges.

The paper introduces a solution called Real-time Linked Dataspace (RLD). It explores the architecture of this solution to demonstrate how it can satisfy the requirements of a smart environment. It discusses its 'Pay-As-You-Go' approach for scalability in data management. It also highlights a query feature for retrieving and viewing live and historical data in a combined output. The paper focuses this solution on two layers of the environment: middleware and application layer.

The technical contributions of the paper include the exploration of the RLD solution across five pilot smart environments among industries of energy and water. It lays out the requirements and high-fidelity design of the technique. It presents a pay-as-you-go approach for incremental tier upgrades. It also defines the workings of the query service for the dataspace. Finally, the author follows the OODA (Observe, Orient, Decide, and Act) Loop in implementation of this solution.

II. TECHNIQUES

The techniques introduced in this paper are called Real-time Linked Dataspace.

Realtime Linked Dataspace

A. Dataspaces

A dataspace is essentially an architecture that supports data sources of various formats and schemas to exist simultaneously. It allows the data to exist in its original format, keeping the source context, and thereby enables a loosely coupled set of data sources to co-exist. It also removes the need to pre-process the data by delaying the time and compute intensive tasks until they are required by some query. When data from two disparate data sources is needed it performs a mapping of sources into a combined schema for integration of data. Therefore, this need-based data transformation makes the process much faster and requires much less storage. This is the main difference from a typical data

warehousing solution that performs data integration on-the-fly. It comes along with a Dataspace Support Platform to aid in its function.

B. Dataspace Support Platform (DSSP)

This is a platform within the dataspace that contains joint support services needed by the data sources for e.g., keyword research. It provides a base implementation for integration of data sources to ease the developers from common data specific tasks and allow focus on application issues. Initially, data sources are loosely coupled and exist on their native infrastructures. However, in case of pay-as-you-go, it must provide the necessary tools to support this tighter coupling. Despite all the tools, the accuracy of its estimations depends on the quality of the data available at the time of querying.

C. Architecture

The architecture of an RLD includes all the physical components such as things, sensors, storages as well as the software data, tools and displays. It must manage the relationships between these items and handle information coordination from diverse sources. The RLD extends its innovation beyond others by also providing real-time querying and processing capabilities for not only the data sources but also their entities. Namely, the architecture consists of the following:

(a) Things/Sensors: Physical items that produce streams of data in real-time. This data is later needed for processing and management. In a smart environment these can include connected mobile/Bluetooth devices, electrical sockets/appliances, water and temperature sensors, and energy utilizations.

(b) Datasets: Heterogeneous information from the things/sensors in a variety of raw formats. Each format is accessible through a different interface of the system. It provides context about the item in the environment. For e.g. In a smart environment these can include water and energy consumptions, passengers flight data, financial information, weather forecasts, and open datasets. They can be in various formats like csv, txt, raw text blobs, numbers, binary etc.

(c) Managed Entities: They are virtual instances of real-world data-emitting items. They contain meta-data about the item such as type, identity, connected data sources etc. They are connected with their relevant datasets and entities to mimic the actual world into a smart environment. They are managed in connection to maintain the inflow/outflow of relevant data on-need basis such as querying.

(d) Support Platform: Contains tools for managing integration of data in the environments. Provides the base functionalities that are essential for developers. Key components are the (1) catalog service, and (2) the real time entity-centric query service. It helps to support scalability of the platform.

(e) Users, Apps, and Analytics: These are software that utilize the data from the platform for analysis, visualizations, decision support tools and user interface dashboards. They allow handling of data access to users/apps with the help of the Human Task service. For e.g., mobile applications, web applications and business intelligence dashboards.

D. Catalog service

It provides the participants of the environment to have access to features such as browsing, searching, and querying the entities to monitor the processes in a manner they wish. It manages the datasets, users, apps, and managed entities in such a way that it can extend the functionality of the original portal. It supports different tiers of service levels each with a different extent of functionality:

(a) Registry: This is a simple register of entities where each of them points to an interface that displays its connected dataset.

(b) Metadata: Additional data about the physical items, coupled together with their emitted data to complete the digital description of entities. Data streams are stored in schema formats. A non-machine-readable format is common for entities e.g., PDF.

(c) Machine-readable: To facilitate querying the metadata needs to be translated into machine-readable grammar and that is done with the help of mappings that convert metadata into query friendly format.

(d) Relationships: The relationships between entities, concepts, and schemas across the smart environment dataspace.

(e) Semantic Mapping: The data from various formats needs to be compatible not only in syntax but also logic. These semantic mappings help manage relationships among different dataset domains. Therefore, it starts to support reasoning decisions and schema independent queries that filter all kinds of data.

E. Access control service

The acce-ss control service manages se-cure data access via define-d roles. It mediates inte-ractions, verifying access through the catalog. It provide-s varied support levels, from datase-t level access to e-ntity-level control and data anonymization for privacy.

F. Search and query service

The se-arch and query service assists de-velopers and users in finding datase-ts within the dataspace. Users can navigate- by entities or perform a se-arch. The challenge lies in balancing expressivity and usability. The- service offers browsing, basic ke-yword search, structured querie-s using SPARQL, and a natural language interface.

G. Entity management service

In a smart environme-nt, effective de-cision-making applications rely on accurate critical entity information. Tre-ating entities as first-class citizens is a vital aspe-ct of managing information in this context. It contains information about the entities in the form of URIs where more data can be retrieved about the entity and its usage across the environment.

H. Entity centric real time query service

In a smart environme-nt, it is vital to facilitate real-time data stre-am queries. The RLD e-mploys an entity-centric real-time- query service allowing unifie-d queries across live, historical stre-ams, and entity data for comprehensive- entity-centric views. It is implemented in RLDSP by following the Lambda Archite-cture.

I. Human task service

The RLD ope-rates the Human Task service-, distributing small tasks among users in the smart environme-nt, promoting collaborative data management. This facilitates user trust and ownership. The se-rvice involves task assignment and quality assurance-, supporting various levels, from none to data quality e-nhancement tasks. Apps within the RLD or its support se-rvices can activate this service-. For instance, users may enrich critical e-ntities using human tasks for entity enrichme-nt based on their environme-ntal knowledge.

J. Five stars pay-as-you-go data management.

The RLD ope-rates on a pay-as-you-go model for managing data, minimizing initial expe-nses. Publishers cover the- joining costs, enabling flexible growth with participants entering, or leaving at any time. Data manage-ment is tiered, whe-re active manageme-nt entails higher costs, lowering e-ntry barriers. Tiers are de-fined using a modified 5-star scheme-, considering data source integration with RLD support se-rvices. Integration with RLD can be gradually improve-d over time.

III. CRITICAL ANALYSIS

In the re-alm of emerging smart environme-nts, managing complex data poses a significant challenge-. This article outlines the Re-al time Linked Dataspace (RLD) as a solution, me-rging dataspaces' pay-as-you-go approach and linked data with real-time- query capabilities. Validated in five- real-world smart environment de-ployments, the RLD effe-ctively supports smart application developme-nt and decision support analytics. Its real-time que-ry service mee-ts smart environments' interactive- query latency require-ments.

The solution proposed by the paper has room for further research. Future e-fforts can aim to scale deployments, cut ope-rational costs, improve support services, imple-ment privacy-by-design for personal data, scale- entity management, and back multime-dia data-. One limitation that is not mentioned could be the dependency on the reliability and accuracy of the data sources. When the-se sources delive-r inaccurate, incomplete, or outdated information, it can affect the- performance of the RLD and its outcome-s' quality. This can be especially challenging in real time situations whe-re timely and accurate data is key.

Overall, the paper does a decent job of introducing an effective technique for data management in a smart environment. It hovers on the most common challenges to such a solution and provides adequate experimentation and data to address these. However, this technique is just one of the many that could be developed under the scope of Dataspaces, and they will only continue to get better as increasingly of them come out.