A Hybrid Storage Model for Web Information Systems[*]

Mark Roantree

Interoperable Systems Group, Dublin City University,
Glasnevin, Dublin 9, Ireland
mark@computing.dcu.ie

Abstract. It is often a requirement of Web Information systems that they incorporate data from multiple and heterogeneous data sources. By definition, this means that data for the same object may be spread across different databases and be stored in different formats. In sport science applications, this problem is exacerbated by the fact that data generated through experiments or sensor devices may lack normalisation and synchronisation. The application of a DataSpace architecture and processes provides a landscape in which solutions to these problems can be built. In this paper, we discuss a real-world sport science application where data is continually recorded and stored in different databases types. The different layers in the DataSpace architecture together with the cross-cutting services are exposed to illustrate how these issues are resolved.

Keywords: DataSpace, Web applications, multi-model storage, metadata

1 Introduction

The HealthSense project is a collaboration between the Interoperable Systems Group and the School of Health and Human Performance at Dublin City University in Ireland. The broad aim of the collaboration is the management of all data generated from healthcare laboratory and field tests, with special emphasis on the performance and welfare of high performance athletes. Both research activities and engineering decisions are driven by the needs of the health specialists who provide end user testing and feedback. This paper focuses on the architecture and services specified and built for the Web Information System currently in use. Specifically, we focus on the heterogeneous information sources used to capture scientific data, and briefly discuss the views that integrate these systems.

Previously, the evaluation of physiological responses during both competition and exercise was limited to a controlled laboratory environment, an approach that lacked ecological validity. It was recognised that a more holistic approach, integrated data or evidence from multiple sources was required to improve decision making. Integrating multiple sources of data can be used:

[*] Funded by Enterprise Ireland Grant No. CFTD/07/201

- To view the average heart rate at specified intervals during exercise or competitive games. This is currently a challenge as the synchronisation of team based data is often difficult [1] and sensor data requires integration with player/patient demographics to include a player's maximal heart rate.
- To compare heart rate responses over historical data or across a number of games played in different conditions.
- To determine the effect of environmental conditions on heart rate responses.
- To measure how changes in the size or dimensions of the playing surface and/or the number of players per team can alter the physiological load (determined by heart rate) [2].
- to determine the period of time that individual players exercise above or below a predefined heart rate, or percentage of their maximal attainable heart rate (heart rate max).
- To combine this information with laboratory based data to determine the percentage of maximal aerobic capacity (VO2max), and the percentage heart rate corresponding to the ventilatory and/or lactate threshold.

The above requirements represent offline analysis of data generated from a number of sensor streams and in some cases, require innovative processes to provide solutions. A further problem arises when the specialist requires immediate or live querying of data as it is generated from sensor devices. Providing real-time query responses to coaches as to when a player reaches a pre-determined excessive level of physiological response (e.g. heart rate), either momentarily or for a pre-defined extended time period, has been shown to optimise a player's performance [3]. The coach may also wish to integrate corresponding video and movement (velocity) data to identify the tasks the athlete was undertaking to determine if it was an appropriate physiological response. This information could be used by the coach to design and implement an appropriate intervention. For example, an immediate response may be to remove the player from the current activity, or change the tactic or drill. The ability to query and respond to the information in real-time will allow teams and individual athletes to gain feedback from experts immediately. Furthermore, by providing this functionality in the form of web-based systems, it often eliminates the need to travel to the training site.

1.1 Contribution and Structure

The motivation for this research project is to provide a data infrastructure and query management system for the diverse requirements of sport scientists in a scalable sensor network environment. A principle challenge arises from the fact that we are dealing with real world requirements, that take their data from multiple sources, and that much of this data is in a raw and unstructured format. This presents the first problem: queries are not possible without building low level software to generate predefined queries. This solution has little innovation and is not practical for users. Further challenges arise as queries across multiple players and teams require that data *must* be synchronised across all players to ensure that players are measured using the same criteria. Furthermore, by providing one set of solutions to the above problems, we create new problems. For example, the use of XML to enrich raw sensor data introduces the problem of query performance with large XML caches. Finally, the modern landscape for personal health systems, especially for high performance athletes, requires an approach to manage the *data everywhere* scenario that now presides. The distributed or federated database architectures will not suffice here as the levels of autonomy and degrees of heterogeneity across systems are too great.

Thus, the goal of this project (the focus of this paper) is to provide an architecture that is flexible enough to accommodate the user's hybrid storage requirements; incorporate the services required to provide base functionality of queries, integration, updates; and finally to provide autonomy for key systems and certain research and development work packages, while on the other hand, providing integration solutions for processes running over multiple data sources. The adoption of a web-based information model provided the standard interface (across multiple platforms) but a novel method for managing and analysing data sources was required. The concept of the DataSpace system was introduced in [4] and [5] as a solution to organisations such as healthcare or sport scientists who have a requirement for "large numbers of diverse but interrelated data sources". While the adoption of this architecture has been slow (see Sect. 4), we adopt the basic principle and extend it to provide more detail of how it can be effective in managing data from heterogeneous and non-traditional data sources.

The contribution of this paper is the application of the DataSpace architecture to a personal health sensor network. Unlike other sensor network applications, this research provides a holistic approach by exploiting the DataSpace architecture to incorporate data from multiple sources and thus, providing a meaningful query engine on raw sensor data. By incorporating data from demographic databases, it is possible to run detailed analysis on sensor streams on both real-time data streams and offline databases. While the main focus of this work is on describing the WIS and Dataspace architectures, we also provide a brief demonstration of our work in combining multiple sensor streams through the functionality of our Metadata Service.

The paper is structured as follows: in Sect. 2, we describe the DataSpace architecture as it provides for a pHealth Sensor Network; in Sect. 3, we describe a sample query with some of the times recorded; while in Sect. 4 we discuss similar approaches; and in Sect. 5, we provide an overall summary.

2 Sensor Web Architecture

In this section we describe the major components of the HealthSense DataSpace system and how each component contributes to the information management process.

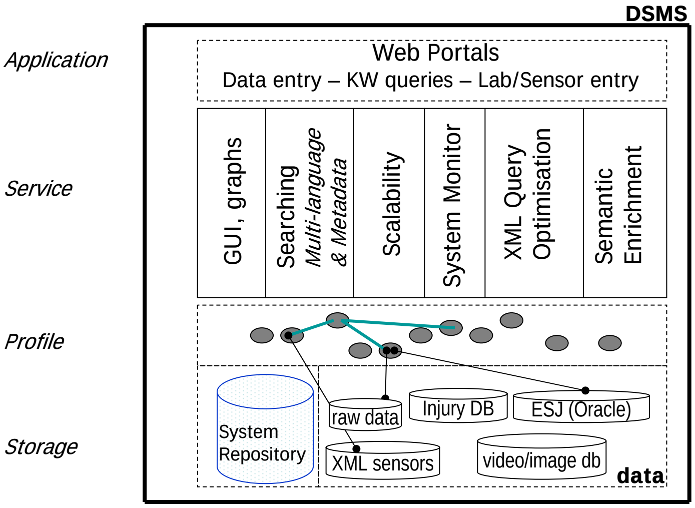

Fig. 1. SportSense DataSpace Architecture

2.1 Storage Component

The HealthSense Storage Component comprises both data sources and a metabase that is used for understanding the content and semantics of the actual data sources. The key difference between the DataSpace architecture and more traditional distributed architectures is that the data is *subject* oriented, similar to a Data Warehouse. In this DataSpace, sensor data exists in both raw format (binary or textual files) and in an enriched XML format. The raw format is necessary for live queries as the converted files are not available quickly enough for this purpose [6]. An Object-relational model (Oracle 10g) is used for creating subject (athlete or patient) profiles and is also used for managing injury data. A separate system is used to store video data while the video metadata is stored in a relational database. Unlike federated database systems that are created from existing application databases, this is a green-field architecture with some level of control to make integration of data easier. However, there will always be situations where unplanned integration will be necessary. For example, combining the output from new sensor devices with previously generated data.

DataSpace Repository The repository for the HealthSense DataSpace provides the engine for the DataSpace system and has a complex metamodel. As with the DataSpace System itself, the System Repository (or metabase) also adopts a hybrid storage model structure. This is necessary as some of the constructs involved are not suited to traditional storage systems. While a full description of the repository and its Metadata Service form part of a separate body of work [6], we provide a brief description of the major components now.

- **Integrations**. Relationships across separate data sources with semantics for integration.
- **Profiles**. For each user or user type, a profile is created that links the user to specified data sources. It may provide a link to Integration objects where users are managing data from multiple sources.
- **Templates**. Templates are used to describe raw sensor sources. Together with a structural enrichment process, they form XML schemas and are used to populate these schemas with raw data so that they can be queried using XPath or XQuery [1].
- **Contexts**. While template objects provide for the creation of XML data from raw sensor output, this provides only a structural enrichment of the sensor data. With Context objects, it is possible to semantically enrich the file. Contexts provide the necessary background to understand the situation in which each sensor was used. For example, a Heart Rate monitor can be used in a match situation: football, tennis or athletics; or it may be used in testing scenarios such as the Bangsbo test [7].
- **Schemas**. Schema objects are stored for XML databases only. They provide the user with a storage model version of raw sensor data and enable the user to formulate queries. There is a strict one-to-one mapping between templates and schemas.
- **Replicas**. There are many examples of data replication in the DataSpace system and these are modelled in the system repository. For example, the optimiser will create a relational index of an XML database; multimedia metadata is created for video files; XML views are created from object-relational databases for the purpose of sharing data.

2.2 Profile Component

HealthSense is a Web Information System in that web browsers provide the interface to multiple sources of data, and HTML or XML is used as an interface between heterogeneous data collections and users. What connects user types with different requirements, to one or more data sources are the profiles. Many profiles are simple to construct: a physiotherapist will only query and manage data from a single database (Injury DB) or a knowledge worker who wants to average team heart rate data after each match. However, some of the profiles are more complex: knowledge workers wishing to mine larger data volumes are searching for relationships between training regimes (Electronic Sports journal), maximal heart rate and injuries sustained.

2.3 Service Component

All of the Service Components represent research and development projects. For the HealthSense DataSpace, the XML Optimisation Service has been created but continues to form ongoing research [8]. The Semantic Enrichment service has been completed for offline sensor analysis [9, 1] but is part of current research for querying live streams.

Query Service. There are two forms of standard query interface: XQuery and a hybrid of XPath and SQL, both accessible through standard web browsers. Unlike all of the existing sensor networks we have encountered, we provide database-style query functionality. This is achieved by the Semantic Enrichment service that delivers XML versions of all sensor data streams. Together with the Metadata Service, a sensor stream is attached to an individual who is described in the Electronic Sports Journal (ESJ) database as seen in figure 1. In other words, as soon as new sensor data is received, it is ready for processing within a short period of time. The enrichment effort is described in [9], together with the times required to both structurally edit and semantically enrich sensor data sources (64MB in less than 1 minute).

Semantic Enrichment. This service creates the XML version for all sensor streams using a template approach [9] that requires no modification to system components when new sensors are introduced. Initially, raw sensor streams are structurally enhanced to produce basic XML files and subsequently, these files are mined to generate additional semantics for every sensor reading. In all sporting experiments, we apply *state* information to each sensor reading, where a state refers to some interval in either a sporting event (eg. football game or tennis match) or a lab-based training activity.

2.4 Application Component

The application layer is deliberately kept simple and serves as a portal to underlying data sources. Thin web browser applications are used to query, run update operations (for example recalibrate sensor readings), run analytical functions and view video data.

3 Case Study and Experiments

On an ongoing basis, data is collected from a series of experiments conducted on teams playing Gaelic (Irish) football, practice tennis matches, and laboratory tests. The application area for this research is typical of a heterogeneous data environment and well-suited to a DataSpace architectural model with sensor data collected both wirelessly and through USB connections. Some of the existing sensors are now described:

- **Polar S625X™ heart-rate monitor**. This consists of a fabric band which fits around a person's chest and generates heart rate data every 5 seconds. It also includes a foot pod that accurately captures velocity and the distance covered.
- **BodyMedia SenseWear®**. This sensor array is worn around the upper arm and measures and logs the following: *galvanic skin response*, a measure of skin conductivity which is affected by perspiration; *skin temperature*, which is linearly reflective of the body's core temperature activities; *heat flux* which

is the rate of heat being dissipated by the body; *subject motion* using an in-built accelerometer. Data values are generated every 60 seconds.
– **iPod Nano 4G with Nike®+ IPod Sport kit**. This sensor includes a foot pod and a detector connected to the IPod Nano. Distance covered during a walk or run together with caloric consumption is recorded.

3.1 Data Extraction

The application layer contains a series of web-based applications for player diary, sensor, video and injury data. The sensor data is uploaded by players for individual data, and by technicians for team-based or laboratory-based data. In figure 1, sensor uploads will impact the raw data and XML sensor databases, together with the relational index for all XML schemas.

Example 1. BodyMedia Data

```
<user>
 <id>1</id>
 <session>
  <id>1</id>
  <sensorData deviceID="bsd">
   <startTime>1195226100000</startTime>
   <measurement time = "1195226160000"
         type="BodyMediaSenseWearData">
    <skin_temp_average_original_rate>30.126686096191406
    </skin_temp_average_original_rate>
    <energy_expenditure_per_minute>1.4301998615264893
    </energy_expenditure_per_minute>
   </measurement>
  </sensorData>
 </session>
</user>
```

A sample BodyMedia enriched to XML format is illustrated in Example 1. Each sensor device has a small XML template file associated with it. The enrichment processor combines the template and output from the device to create an XML schema that can be queried using the XPath query language. If the system encounters a new sensor, all that is required is its template. The template describes the sensor's output in terms of its structure and the location of key elements, such as start time and measurements, as well as important data such as value delimiters. The output file contains user, session and device ID information, followed by sensor-specific information and finally, a list of labelled measurement values. A more complete discussion on this process is provided in earlier work [9].

3.2 Data Enrichment and Integration

Any number of sensor streams may be merged in advance of user queries. As sensor data is not as complex as database schemas, the integration process is not at the same level of difficulty as in federated or mediated schemas. However, this is replaced with issues such as normalisation of data and synchronisation across teams and multiple experiments [1]. The DataSpace Repository nominates various attributes that are used as axis points for integration eg. time as shown in Example 2, where data is merged from heart rate and BodyMedia sensors.

Example 2. Integrated Data

```
<user>
 <id>1</id>
 <session>
  <id>1</id>
  <sensorData
   <measurement time="1195226130000">
    <HRValue>91</HRValue>
   </measurement>
   <measurement time ="1195226160000">
    <skin_temp_average_original_rate>30.126686096191406
    </skin_temp_average_original_rate>
    <energy_expenditure_per_minute>1.4301998615264893
    </energy_expenditure_per_minute>
    <HRValue>95</HRValue>
   </measurement>
   <measurement time="1195226190000">
    <HRValue>97</HRValue>
   </measurement>
  </sensorData>
 </session>
</user>
```

The queries shown in Table 1 are examples of simple XPath queries that can be used when data is required purely from XML sources. In many queries, it has been necessary to construct a hybrid of XPath,SQL and Java to extract the information described in the introduction to this paper. Space issues prevent a deeper discussion of how these queries are constructed but the purpose of this section is to demonstrate how raw sensor data can be queried in the existing DataSpace architecture.

XPath Expression	Query
//user[id="Sarah"]//measurement [skin_temp_average_original_rate [text() >30]]	Return all sensor readings for Sarah when her skin temperature is higher than 30
//user[id="Fionnula"]//measurement [HRValue[text() >120]]	Return sensor readings for Fionnula when the heart rate exceeds 120
//user[id="Aoife"]//measurement [skin_temp_average_original_rate[text() >30] and HRValue[text() >130]]	Return sensor readings for Aoife where the skin temperature exceeds 30 degrees and heart rate exceeds 130

Table 1. XPath Query Examples

4 Related Approaches

Although DataSpace systems were first introduced in 2005, research is this area is not widespread and in many cases, it is used to address the financial side of information systems where access to information and services requires payment. While we do not currently address the pricing aspect of information access, other characteristics of these systems are common between our approach and that of others.

Research work that predates the DataSpace approach but is very similar in requirement to our work can be found in [10]. In this work, the authors also claim to provide an automated integration into the scientific database, and their architecture has a similar structure to HealthSense. While they employ a method of algorithmic analysis to process incoming data on an automatic basis, we employ a template-based process to ensure that all data streams are properly processed and transformed. Thus, we can greatly reduce the number of errors when importing data. Any issues to do with integration are deferred until query processing. The other major difference is that they employ relational database technology for data storage while we do not believe that this approach can adequately manage the heterogeneous data types that are generated in the sports science domain.

In [11], the authors describe a problem domain similar to the HealthSense project in that users require access to, and management of data located in both structured and unstructured sources. This work provides more detail than we can provide in this paper on the hybrid query language required for heterogeneous data sources. The work is heavily focused on a keyword approach to data retrieval only. In [12], the same research group describe a process for managing uncertainty across multiple relational databases. While this provides a significant advance on integrating data sources with uncertainty and is focused on the DataSpace environment, it does not address the larger issues of DataSpace Management, the required services, or the continuous importation of sensor data in raw format.

In [13], the authors present a Personal DataSpace Management System called iMeMex. This is similar to our work in that it provides a holistic approach to data management: files, emails, pictures, videos and calendar are all part of the system's repository. They employ a DataSpace platform and also similar to our approach, they provide an operational prototype. However, their focus is limited to data for a single individual whereas we are required to manage data for multiple users from heterogeneous data sources and to manage the data required by sensor devices.

5 Conclusions

In this paper, we describe a collaboration between a data engineering research team and a group of well established sports scientists working with both elite sports men and women, and members of the general public undergoing health trials. As is typical with many scientific data collections, it was quickly identified that a single point of storage, with a single data model would not provide a solution to their needs. A multi-storage system with standard access facilities emerged as key requirements. The adoption of a web-based information model provided the standard interface (across multiple platforms) but a novel method for managing and analysing data sources was required. The HealthSense System is a fully operational prototype that works with real-world data and as a characteristic of a DataSpace architecture, it permits ongoing research into some services while users can still import sensor data and run daily queries.

Our current efforts are focused on optimisation of both database and streaming data. On the database side, we are continually optimising both read and update operations on XML databases, while also building a query service across all hybrid data sources. For querying live sensor data (during match time or training activities), XML streaming approaches cannot be adopted as the time required to convert raw data to XML is not acceptable to health specialists. Thus, a hybrid streaming approach has been developed [6] to use schema information from the metabase to map XPath queries to lower level querying constructs.

References

1. Roantree, M., McCann, D., Moyna, N.: Integrating Sensor streams in pHealth Networks. In: 14th International Conference on Parallel and Distributed Systems, pp. 320-327, IEEE Computer Society Press (2008)
2. Hill-Haas, S., Coutts, A., Rowsell, G., Dawson, B.: Variability of Acute Physiological Responses and Performance Profiles of Youth Soccer Players in Small-sided Games. Journal of Scientific Medical Sport, 11(5):487-90 (2008)
3. Impellizzeri, F.M., Marcora, S.M., Castagna, C, Reilly, T., Sassi, A., Iaia, F.M., Rampinini, E.: Physiological and performance effects of generic versus specific aerobic training in soccer players. Int. J. Sports Medicine, 27(6):483-92 (2006)
4. Franklin, M., Halevy, A., Maier, D.: From Databases to Dataspaces: A New Abstraction for Information Management. SIGMOD Record, 34(4), pp. 27-33 (2005)
5. Halevy, A., Franklin, M., Maier, D.: Principles of Dataspace Systems. In: 25th Symposium on Principles of Database Systems, pp. 1-9. ACM Press (2006)
6. McCann, D., Roantree, M.: A Hybrid Query Service for Optimising Raw Sensor Data. Submitted for publication (2009)
7. Chamari K. et al.: Field and Laboratory Testing in Young Elite Soccer Players. British Journal on Sports Medicine, vol. 38, pp. 191-196 (2004)
8. Marks, G., Roantree, M.: Pattern Based Processing of XPath Queries. In: 12th International Database Engineering and Applications Symposium, pp. 179-188 (2008)
9. Camous, F., McCann, D., Roantree M.: Capturing Personal Health Data from Wearable Sensors. In: International Symposium on Applications and the Internet (SAINT), pp. 153-156. IEEE Computer Society Press (2008)
10. Leser, U., Naumann, F.: (Almost) Hands-Off Information Integration for the Life Sciences. In: 2nd Innovative Data Systems Research (CIDR), (Online www-db.cs.wisc.edu/cidr) pp. 131-143 (2005)
11. Liu, J., Dong, X., Halevy, A.: Answering Structured Queries on Unstructured Data. In: 9th International Workshop on Web and Databases, pp. 25-30. ACM Press (2006)
12. Dong X., Halevy A., Yu C.: Data Integration with Uncertainty. In: ACM 33rd Very Large Databases, pp. 687-698. ACM Press (2007)
13. Blunschi L. at al.: A Dataspace Odyssey: The iMeMex Personal Dataspace Management System. In: 3rd Conference on Innovative Data Systems Research (CIDR), pp. 114-119 (2007)