Information Fusion Framework (IFF): A deploy for the electricity market domain

Paulo Trigo and Paulo Marques

***Abstract*—**The Information Fusion Framework (IFF) unifies dissimilar information sources into a single and consistent model. Thus, IFF frees the users from tasks such as finding the relevant information sources, interacting with each source in isolation and selecting, cleaning and combining data from multiple sources. The IFF unification process enables the (expert) user to specify all the integrity constraint validations (e.g., structural, consistency and referential) to be satisfied regarding each information source. The unified model evolves according to a transactional approach and users are notified (via email) upon transaction termination. Users are also considered as information sources so they may freely provide new or corrective information. The IFF system is currently deployed and configured in the Electricity Market domain.

***Keywords*—**Information Fusion, Data Integration, Electricity Market.

I. Introduction

THE information fusion goal is to combine different sources of information to produce new data that is expected to be more informative and synthetic than the inputs. A simple illustration of fusion is the use of various sensors to detect a target or to build a landscape image. Another example is the exploiting of the media news (e.g., newspaper, radio, internet and television) to support a stock-market biding decision.

Additionally, the information fusion process is usually designed to provide a bidirectional mapping from the diverse data sources into an unified view (uniV) and then, from such unified view, back to the original data sources. The construction of the unified view is often regarded as an integration process (i.e., data-to-uniV). Then, the usage of the unified view as an input to the evolution of the data sources is often described as a dissemination process (i.e., uniV-to-data).

This paper presents the Information Fusion Framework (IFF) that provides an environment for the integration (i.e., data-to-uniV) and dissemination (i.e., uniV-to-data(of information. The IFF comprises four major aspects of information fusion, namely:

i. formal description of the diverse concepts, its attributes and the relationships between concepts,
ii. validation of the data integrity constraint requirements for an unified view,
iii. transactional approach to the model's instance evolution,
iv. user notification and participation in the whole process.

The information fusion is useful for several tasks such as decision-making, processual workflow and target tracking. Those tasks are found in many application areas such as Market Analysis, Logistics, Defense, Robotics and Medicine, among others. The general idea is that the information fusion process improves the knowledge retrieval capabilities of the overall resulting system.

The IFF system is being developed in the context of the Electricity Market domain and we intend to further extend its functionality and to adapt its general-purpose information fusion support to other application areas.

The implementation of the IFF (kernel) process resorts to the Java programming language. All the IFF configurable aspects are declaratively specified using the Extensible Markup Language (XML) related technologies, such as the XML Schema Definition (XSD) for structural specification and validation, the XPath language to navigate in XML documents and the Extensible Stylesheet Language Transformation (XSLT) to produce the set of constraint violations of a XML document; the XSLT is also used to describe the data model transformations, e.g., from the XML (hierarchical) model to the relational (flat) model.

Section 2 describes the essential IFF modeling decisions and Section 3 describes the initial feedback from an operational IFF deployment. Section 4 outlines conclusions and future goals.

II. The IFF Modeling

This section describes the IFF support for the integration process, i.e., from the diverse data sources into an unified model. The reverse pathway (i.e., dissemination process) will not be describes in this paper (still not fully supported in IFF).

The IFF construction follows a model-driven architectural (MDA) approach and all the platform independent models (PIMs) are formally specified within the Unified Modeling Language (UML) framework [14]. The platform specific models (PSMs) are built using the appropriate "UML-to-Java" transformations. The IFF supports declarative configuration notion as all the integrity constraints are tailored to each domain as rules (to be satisfied) are formally described using the XML-related technologies and Structured Query Language

Paulo Trigo; Instituto Superior de Engenharia de Lisboa – ISEL; Área Dep. de Engenharia de Electrónica e Telecomunicações e de Computadores – ADEETC; Lisboa; Portugal (e-mail: ptrigo@deetc.isel.ipl.pt).

Paulo Marques; LinkPoint – Lisbon Technology Connection; Lisboa; Portugal (e-mail: paulo.marques@linkpoint.pt).

(SQL) statements.

A. The IFF business model

An UML use case diagram provides the IFF business model overview. The use case diagram is a behavioral model that shows what system functions (or services) are performed for which external entities. System functions (or services) are represented as use cases that are performed to serve the system's external entities represented as actor.

The UML graphical representation depicts a service provider as being pointed by an arrow; the service consumer is found on the other end of that same arrow. Fig. 1 shows the IFF use case diagram. The "Transactional IIntegration" use case (service provider) provides the transactional integration of the information contained in the "XML IStore" actor which may be materialized as either an "Application", an "User" or a "DataBase" meaning that any such actor may be regarded as a source for the information fusion process.

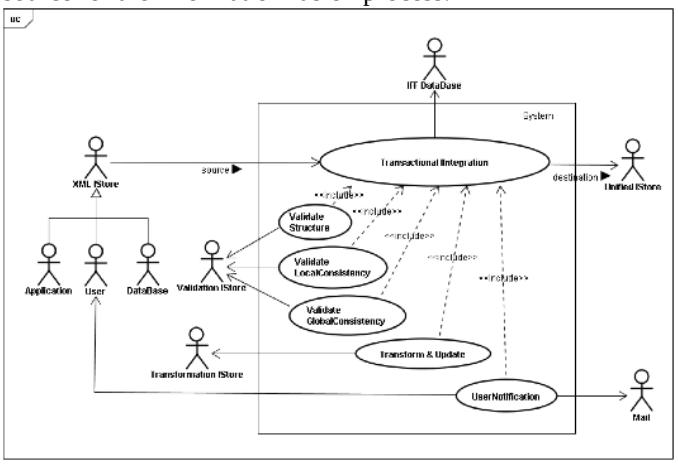

Fig. 1. The information integration business model.

Fig. 1 shows that the service provider perspective of the "Transactional IIntegration" use case resorts (in its turn) to the services provided by the following two actors:

- "**Unified IStore**", actor that supports both the schema of the unified data model and the integrated data instances; actor is currently deployed as an Oracle relational data base.
- "**IFF DataBase**", actor that contains all the IFF specific and configurable data and that also provides the typical short-term transactional service that is used, by the IFF, to implement its long-transactional model.

Additionally, the "Transactional IIntegration" achieves its goal by including a set of validation use cases ("Validate Structure", "Validate LocalConsistency" and "Validate GlobalConsistency") as well as a "Transform&Update" and a "UserNotification" use case. We briefly describe each of those use case as follows:

- "**Validate Structure**", picks each XML document that contains a set of related concepts (and respective attributes) being integrated and performs a "XML × XSD" confrontation.
- "**Validate LocalConsistency**", applies a set of XSLT rules that transforms each XML document into a list of constraint violation detailed messages; an empty list represents a valid XML document.
- "**Validate GlobalConsistency**", incorporates the data of each XML document in the unified model while performing the set of specified validations between each new piece of information and the already existing data in the unified model.
- "**Transform&Update**", applies a set of XSLT rules that transforms each XML document into statements that appropriately update the "Unified IStore", currently, XSLT is used to transform XML documents into SQL statements.
- "**UserNotification**", gives each user the relevant information, according to a configurable user/message profile, about the integration process.

B. The IFF architecture

An UML component diagram gives an overview for the IFF process workflow. The main purpose of the component diagram is to show the structural relationships between the components of a system. Components are considered autonomous and encapsulated units of a system (or subsystem) that provide one or more interfaces; thus being easily reused or substituted by any implementation that satisfies the component interfaces.

The component diagram provides a high-level architectural view of the system and gives (the designer) a formal roadmap for the implementation. It also helps the (project manager) decision-making concerning task assignments and the capabilities to acquire. Fig. 2 shows the IFF architectural perspective.

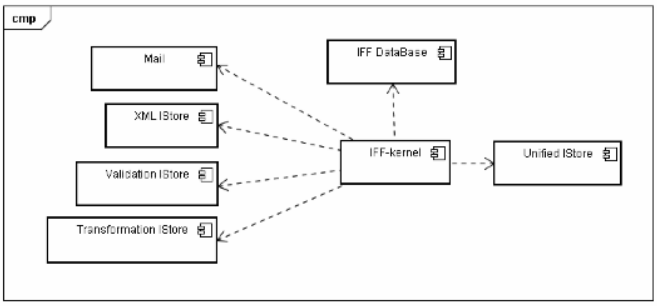

Fig. 2. The IFF architecture (as UML component diagram).

Fig. 2 shows that the "IFF-kernel" component is central to the whole fusion process and that it exploits the interfaces provided by all other components. Apart form the "IFF-kernel" all other components are loosely coupled, meaning that they may be easily substituted or adjusted, e.g., for domain-dependent modeling and configuration purposes.

Additionally, from the Fig. 2 logical architecture several possible physical mappings are available. For example, the current IFF deployment is as follows: a) the "IFF-kernel" is maps to an application server node, b) both the "IFFDataBase" and the "Unified IStore" are mapped into another a data server, c) the "Mail" is provided by a different,

mail server, node, d) all the "XML Store", "Validation DataBase" and "Transformation DataBase" reside in the same IFF node.

Several other "logical to physical mappings" (deployments) are easily devised. Additionally, the overall design may be easily extended by the incremental plug of special purpose components and by accordingly adapting the "IFF-kernel".

C. The IFF process workflow

An UML activity diagram provides the IFF process workflow overview. The activity diagram is a behavioral model typically used for modeling the detailed logic of a business process or to capture the behavior of a single use case or business rule. An activity diagram shows the event(s) that causes an object to be in a specific state and highlights, throughout the swimlanes, the responsibility of each participant (e.g., actor) in generating and managing each event.

Fig. 3 shows the IFF activity diagram and the "IFF-kernel" swimlane evidences the adoption of a state-transition machine approach that takes the contents of each XML document throughout the whole information integration process.

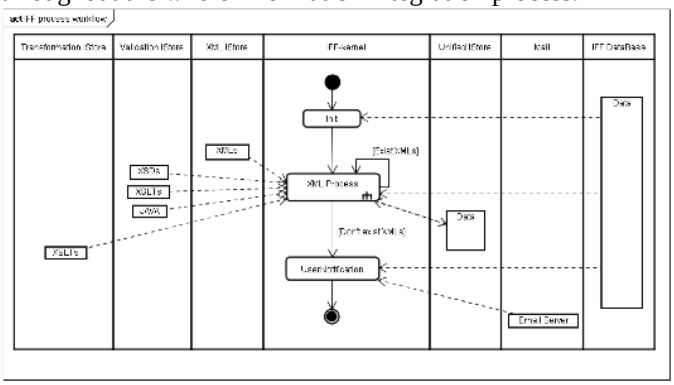

Fig. 3. The IFF process workflow (an UML activity diagram)

The state-transition machine executes as a long-transaction and follows the general idea of the "saga" concept where a long-transaction is built from a pipeline of several typical (short-term) transactions each one with its own compensatory procedure to be executed whenever the related transaction fails. The associated short-term transaction support is taken from an already existing system; i.e., currently, the IFF long-transaction model is implemented over the Oracle short-term transactional service.

D. The IFF-kernel detailed perspective

Fig. 4 exhibits the IFF-kernel details within an activity diagram. This model evidences that the IFF-kernel is designed to support two execution modes:

- A **"daily" mode**. The information integration process is executed once each day within a configurable time slot; before starting the daily operation all the configurable parameters (e.g., the daily execution time slot) are (re)loaded into the IFF-kernel memory.
- An **"intra-daily" mode**. During each day, according to a configurable periodicity, the information integration process is executed; the goal is to incorporate the "ad-hoc" (unplanned) initiatives from the information sources.

The operation of both the "daily" and the "intra-daily" modes only admits the set of previously registered (configured) concepts (contained within XML documents); i.e., only registered XML documents will be accepted and processed.

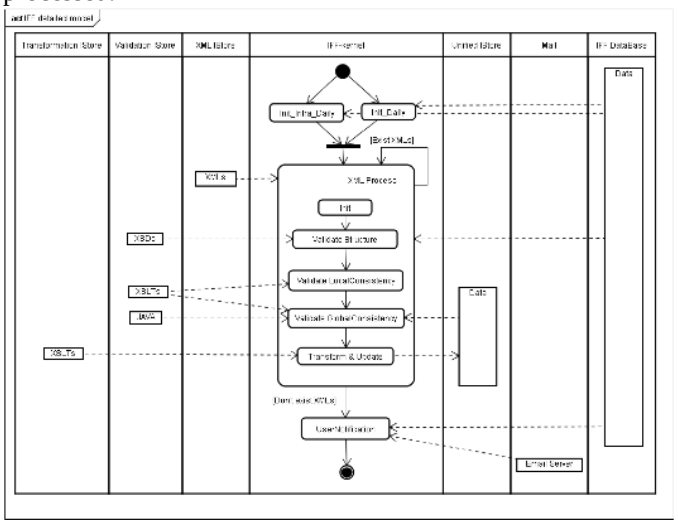

Fig. 4. The IFF-kernel detailed model.

Fig. 4 also shows that the IFF-kernel is designed as an hierarchical state-transition machine, i.e., each state my itself be internally specified as a hierarchical lower-level state-transition machine. Such an uniform formalization of the "IFF-kernel" behavior simplifies the implementation of the long-transaction model and enables explicitly distinguish between different processual concepts such as "executing a XML validation pipeline" or "sending as user notification". The hierarchical modeling also revealed to be useful for the implementation of the fail (network breaks and system crashes) recovery mechanisms.

III. Experiments, Results and Evolution

The IFF current implementation is being deployed and used within the Electricity Market application domain. The deployment has already taken two stages. Each stage was scheduled to take 6 months before moving to the next stage.

The first deployment stage included the "daily" data unification of 35 XML documents with a total information amount of around 5.52 MB. The average IFF file processing time was around 2.5 seconds (a total of 1.28 minutes for the 35 files). During this first stage there were 6 different types of data (i.e., data sources) and 109 sub-types (i.e., different concerns within each data source).

The second deployment stage formally accepted IFF as a sound tool and therefore the sources were extended (from 6) to 19 types of data and 235 sub-types (around 150% work load increase). Also, in this stage IFF deals with several different scheduling periodicities for data integration. Fig. 5 presents a synthetic overview of some of the (most relevant) types of data along with quantitative metrics, such as, the

number of files to process, the overall size of data and the periodicity (h: hourly, d: daily, bd: bi-daily, id: intra-daily and e: episodic), requests by each type of data.

Fig. 5 highlights the following IFF main concerns:

a) simultaneously process a large number of files,
b) process individual files that are moderately large,
c) evaluate a large set of consistency and integrity constraints (not shown in Fig. 5 but there is a large diversity of constraints' evaluation),
d) satisfy the diverse scheduling requirements of each data type.

Also, IFF is now supporting most of the information fusion tasks (previously done with several ad-hoc and distinct tools) and therefore its reliability is becoming a critical aspect.

All the above concerns are taking IFF into an architecture with a high number of IFF-kernel processes executing in different nodes and implementing a fully distributed mechanism for the synchronization of the overall data fusion process. Additionally, IFF is being redesigned as to parallelize its (internal) hierarchical state machine (formulating some states as independent threads) and improve the computing resources' management. The details on IFF distributed and state-thread approaches are still under development and therefore they are beyond the scope of this paper.

Data Source (Type of Data)	#files	size	period.
INTERCONNECTION-CAPACITY	5	1 Mb	e
PARAMETERS	1	25 Kb	bd
HOURLY-MARKET-OFFER	1	3 Kb	h
DAILY-MARKET-COMMUNICATION	—	—	—
PROGRAMS	7	40 Kb	d
INTERCONNECTION-CAPACITY	3	43 Kb	d
NETWORK-POWER-FLOW	1	210 Kb	d
PREDICTION (P) AND REAL (R):	—	—	—
(P) DEMAND	28	90 Mb	h & d
(P) NETWORK-POWER-FLOW	1	8 Kb	h
(P) MARKET-PRICE	1	68 Kb	d
(R) DEMAND	23	300 Kb	d
(R) NETWORK-POWER-FLOW	1	40 Kb	3xid
(R) POWER-GENERATION-DIAGRAM	1	23 Kb	bd
SPOT-MARKET:	—	—	—
PRICES-DAY & PRICES-INTRA-DAY	2	6 Kb	id & d
SPOT-LAST-30-DAYS	7	1.5 Mb	d
SPOT-HISTORY	7	47 Mb	d
FORWARD:	—	—	—
PRICE-POSITIONS	23	340 Kb	d

h: hourly | d: daily | bd: bi-daily | id: intra-day | e: episodic

Fig. 5. Overview of data heterogeneity being unified by IFF.

Currently IFF is also being extended to deal with the integration of the "Gas fuel" in the electricity market analysis, which represents a high increase in the number of the types (and sub-types) of data and also on the scheduling of the several unification processes.

IV. Related Work

The IFF approach crosses the research fields of data-warehousing (DW) [1] and data-mining (DM) [2]. Both DW and DM provide specific interpretations for the "data fusion" concept but, despite the differences, the (common) final goal of data fusion is to integrated and correlated disparate data stores with each other without having to join or rebuild everything, without having to build a huge number of store and data type interfaces, without having to build pattern-specific detection filters or having to build data-model specific storage architectures [5].

The DW perspective on data fusion is closely related to the ETL (extract, transform, load) process that usually aims on the construction of *data-mart(s)* which are "departmental-focused" views of data maintained separately from the organization's operational databases. The DW supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from those of on-line transaction processing (OLTP) applications supported by the operational databases [3]. The IFF purpose is to unify data but keeping the "global database" perspective along with its operational OLTP capabilities. In this respect the IFF can provide the input to a subsequent DW process of *data-mart* construction.

Some research on DW deals with the heterogeneity among different data sources and applies models (and technology) founded in description logics (DL), and its implementation in the Semantic Web context, to abstract local ontologies from metadata of different data sources and then reason on the identified relations to get ETL (prospective) rules that can assist the performing of DW project [4]. This approach is formally sound and we intend to explore it in the IFF context. The derivation of schema equivalence, or the entailment of rules to bridge the semantic gap between schema, is a general problem that follows a long line of research [6], [7], [8]. The IFF takes an operational perspective and (at its current stage) schema equivalence (or mapping) is formulated at design time.

The DM perspective on data fusion is often formulated as a pattern-matching problem where two datasets share a set of variables and the closely related data (i.e., similar data-pattern on those shared variables) is used to predict the value of the non-shared variables [5], [9]. The multi-sensor management [10], [11], is an important application of this data fusion perspective. The DM projects can be very complex and may need to follow a specific process, e.g., the CRISP-DM [13]. The IFF approach follows the general goal of predicting some data given the gathered evidence but such goal is currently explicitly designed by combining shared data (variables) to specify (computed) values of non-shared variables. Also the IFF development and deployment process is less aligned with CRISP-DM and much closer to software engineering methods (e.g., the RUP [15]).

V. Conclusions and Future Work

This paper describes the main modeling aspects of the Information Fusion Framework (IFF) system. The current implementation stage is focused in the information integration

perspective, i.e., automatically build an unified model that comprises several diverse data sources.

The IFF deployment in the electricity market domain was scheduled in two 6-month stages with an increasing demand on the diversity of data sources and on the size of the data to get unified.

The IFF adoption in the electricity market domain enabled to iterate toward a mature solution for data fusion and we are currently aiming to expand and apply the approach to new application domains. Shortly the IFF will also incorporate some information dissemination features.

Therefore, future work regarding the IFF deployment includes exploiting and adapting the IFF to other related application domains.

As for the future research line we are considering to use the multi-agent design paradigm in order to extend the IFF into a distributed environment; e.g., resorting to the Extensible Messaging and Presence Protocol (XMPP) based middleware. Along this line of research we also intent to explore the description logic formalism, which supports, for example, the Web Ontology Language (OWL) specification, as a means to reduce the gap between the user defined and the (machine) autonomous mapping of dissimilar concepts.

REFERENCES

[1] Surajit Chaudhuri and Umeshwar Dayal, "An Overview of Data Warehousing and OLAP Technology", in *Journal of SIGMOD Rec.*, vol. 26, no. 1, pp. 65–74, 1997.

[2] Ming-Syan Chen, Jiawei Han and Philip S. Yu, "Data Mining: an Overview from a Database Perspective", in *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866–883, 1996.

[3] Simitsis, A.; Vassiliadis, P.; Sellis, T.; "State-Space Optimization of ETL Workflows", *in IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no 10, pp. 1404–1419, 2005.

[4] Zhuolun Zhang and Sufen Wang; "A Framework Model Study for Ontology-Driven ETL Processes", *in the 4th International Conference Wireless Communications, Networking and Mobile Computing WiCOM'08*, pp. 1–4, 2008.

[5] John Shawe-Taylor, Tijl De Bie and Nello Cristianini, "Data Mining, Data Fusion, and Information Management", *in the IEE Proceedings - Intelligent Transport Systems*, vol. 153, no 3, pp. 221-229, 2006.

[6] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Modeling ETL Activities as Graphs," in the Fourth International Workshop on Design and Management of Data Warehouses, pp. 52-61, 2002.

[7] S. Alagic and P.A. Bernstein, "A Model Theory for Generic Schema Management," *in the Eighth International Workshop on Database Programming Languages*, pp. 228-246, 2001.

[8] R.J. Miller, Y.E. Ioannidis, and R. Ramakrishnan, "Schema Equivalence in Heterogeneous Systems: Bridging Theory and Practice", *in Information Systems*, vol. 19, no. 1, pp. 3-31, 1994.

[9] Peter Van, Der Puttan, Joost N. Kok and Amar Gupta, "Data Fusion Through Statistical Matching", *in Social Science Research Network Electronic Paper Collection*, 2002.

[10] N. Xiong , P. Svensson, "Multi-sensor Management for Information Fusion: Issues and Approaches", *in Information Fusion – ScienceDirect*, pp. 163–186, vol 3, no 2, 2002.

[11] J Manyika and HF Durrant-Whyte, "Data Fusion and Sensor Management: a Decentralized Information-Theoretic approach", *Ellis Horwood*, 1994.

[12] Rüdiger Wirth and Jochen Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining", *in Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pp. 29-39, 2000.

[13] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, R. Wirth, "The CRISP-DM Process Model", *in Draft Discussion paper Crisp Consortium*, http://www.crisp- dm.org/, 1999.

[14] Ivar Jacobson, Grady Booch and James Rumbaugh, "The Unified Software Development Process", Addison Wesley Longman, 1998.

[15] Philippe Kruchten, "The Rational Unified Process: An Introduction", Pearson Education, 2000.