

# Using a Business Activity Monitoring and SOA for a Real-Time ETL

Abdelmgeid A. Ali  
Faculty of Science,  
Computer Science Department,  
Minia University,  
Minia, Egypt

Waleed M. Mohamed  
Faculty of Computers and Information,  
Computer Science Department,  
Minia University,  
Minia, Egypt

## ABSTRACT

Data that is obtained from various information sources needs excessively handling for managing, analyzing and monitoring. Data warehouse consolidates data coming from different data sources. Data warehousing technology has made a huge effect in the world of business; it transforms information into data that helps analysts to make strategic decisions. A real-time data warehouse is used same purpose as data warehouse, in addition to these, data streams into real time data warehouse on time. In this paper, propose a real-time framework to support the up-to-date process. Our framework is based on using a Business Activity Monitoring (BAM) that provides a real-time business intelligence by capturing data as it flows through a business system. Using service-oriented architecture (SOA) and Windows Communication Foundation (WCF) to capture a real-time data and load easily. By using BAM, it can monitor a business process in real time.

## Keywords

Real-time data warehouse (RTDW), Change Data Capture (CDC), Business Activity Monitoring (BAM), Windows Communication Foundation (WCF).

## 1. INTRODUCTION

The term data warehouse (DW) is commonly used in manufacture and it denotes a kind of heterogeneous data sources. Firstly discovering that a data warehouse is an environment, not a product. The need for building a data warehouse is that corporate data often dispersed in various data sources and possibly in different formats. In order to view a complete picture of information, it is necessary to access these heterogeneous data sources. Therefore, it can be obtained data and pieces of partial information from each data sources, and then put them together to produce an overall picture. Attempting this process without a data warehouse is a hulking task, inefficient, and error-prone. Moreover, this task will usually need huge efforts of system analysts. All these difficulties discourage the effective use of complex, but valuable corporate data [1]. Bill Inmon [2] defines the term DW as: “A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management’s decisions”.

Business time is progressively moving towards real time – so as to grow their competitive advantage, organizations should seek to capture and react to business events faster and more strictly than ever. Today, the prevalence of competitive advantage comes from the effective use of information technology. Therefore from an IT standpoint, the key tool of the trade for enabling business intelligence is the enterprise data warehouse, combined with an enterprise analytics

framework. Traditional data warehouses have been strategic-only resources that help create reports, analyze events, and predict what might happen in the future. Today’s data warehouses are not only strategic but also tactical; adding mission-critical decision support to their workload. Data warehouse can be revived in many different ways. The mechanism of reviving the warehouse is very important in deciding the reliability of warehouse’s content. Traditionally data warehouse were revived at specific intervals, e.g., every night, weekly, or monthly. Due to demand for fresh data in data warehouse we need to shift from traditional data warehouse to real time data warehouse. WCF is a programming model for using managed code to build unified web service and other distributed businesses that can talk to each other and deploying services among various endpoints. Using WCF, it can send data as asynchronous messages from one service endpoint to another. A service endpoint can be part of a continuously available service hosted by internet information services (IIS), or it can be a service hosted in an application. An endpoint can be a client of a service that requests data from a service endpoint. Web Services [3] can be accessed only over HTTP, where WCF is adaptable because its services can be hosted in different types of applications.

This overlooking of BAM is unfortunate, because BAM bridges the gap between business processes and business intelligence to provide rich capabilities for real-time business intelligence. If you have business processes or web services implemented on any an application platform, BAM provides a high-performance infrastructure for monitoring the behavior of your systems and brings a powerful real-time component to any business intelligence strategy.

BAM can capture data from the Simple Object Access Protocol (SOAP) envelope in a WCF request or response message. The captured data is specified using an configuration file that called interceptor configuration (IC) file, an Extensible Markup Language (XML) file that is created by a developer. SOAP allows processes running on various operating systems (such as Windows and Linux) to communicate using XML.

## 2. RELATED WORK

Various works in the area of real-time ETL processes have been published. Obali and etal [11] proposed a RTDW based on the Web services and use Change Data Capture (CDC) technique to get the data changes from raw data systems and send them to the web service. The real-time ETL and data integration are not treated. Lebdaoui and etal [12] treated the problems of data integrity issues towards the requirement for

getting the RTDW and presented an IA-RTDWg model that maintains, at the same time, the data integrity and accessibility. It use CDC technique to get the changed data. They not treat The real-time ETL and not present it's architecture. In the [13] The authors suggested a solution for the near real time storing , with the RTDW architecture from a basic DW architecture. That updates the constant data while preservation the most recent information in DW and enabling any required degree of freshness to queries. They can't treated the real-time ETL. previous work [14] proposed a new model called DJ-DI to treat the problem of the integration big data in the data warehouse in a short time. The authors presented the RTDW architecture to enhancing data integration, while keeping the data integrity respecting the existing safety rules. This model, which is built on partitioning the size of data changes and the duplication of the fact table. The real-time ETL are not treated. In [15], the authors identified the problems in near real-time DW and their solutions at the level of each phase of the ETL process. They identified two problems for each extraction, transformation and loading phase. They not presented the RTDW architecture. Referring to previous related work to the current research, it turns out that includes RTDW architecture as fully integrated real-time ETL.

### 3. SOA AND BUSINESS ACTIVITY MONITORING

The aim of using WCF service, asynchronous communication between a client and a service, which may be on various machines. Using MSMQ (Microsoft Message Queuing) as a transport system. MSMQ will be deployed on a middleware server. The client application will use WCF's *NetMsmqBinding* to send messages to a private queue on the middleware server. The service hosted by IIS, then listen for new messages on the message queue. This listening is accomplished by a Windows Service. At the point when a message arrives hen a message arrives, it invigorate the service within an IIS worker process, and the service will process the message. MSMQ is designed to be resistant against all kinds of failures, such as temporary network outages. Figure 1 shows these operations.

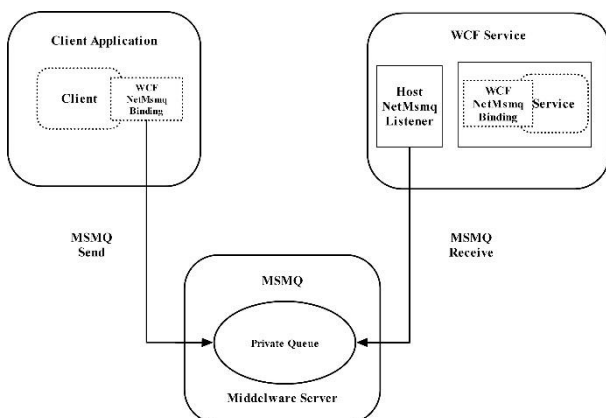


Fig.1: WCF service architecture model.

The goal of BAM is to provide real-time business activity monitoring through computation and visualization of key

performance indicators (KPIs) [4]. presently, BAM is normally integrated as a feature of Enterprise Resource Planning (ERP) systems or workflow management systems. One of the fundamental objectives of BAM is to minimize the latency [5] of decision making by offering support for immediate operational decisions and for taking suitable actions based on real-time information [6]. BAM is frequently used to capture data from applications built using service-oriented architecture principles. Through logical analysis of the BAM framework, the data and functions are encapsulated in objects. These objects perform specific functions through the coordination of communication between each other. Object oriented system model has low overhead, low complexity features, high flexibility. But performance of the system is seriously affected by the number of objects and the consistency of data and its format is not resolved [7]. The rise of SOA development model fetch new opportunities to BAM's development. SOA gives a perfect integrated architecture of reusable services and supports dynamic loading of service components in business flows. However, the existing SOA do have a few imperfections, lacking the abilities of dynamic significance, cooperative working and asynchronous processing. The event-driven architecture can compensate for these SOA troubles [8].

SOA is a style of design in view of message going between systems, where the objective is to expose data and services in a loosely coupled manner that makes it simple to make composite services.

SOA projects typically follow four essentials:

- Boundaries are explicit: Calling a service may require crossing a network, crossing a security boundary, or marshalling data between various representations. Service-oriented designs pay attention to the costs of boundary crossing; messages are passed explicitly instead of being implicitly created by method calls.
- Services are independent: Services are formed and managed autonomously of their clients. Changes in implementation details are irrelevant to the clients.
- Services share schema and contract, not class: In SOA, there's a bright line between the external interfaces to the service and the internal implementation of the service. The interfaces to the service are typically defined using Web Services Description Language (WSDL). Implementation details, such as classes, are hidden from the client. All interaction with the service is through endpoints defined by contracts.

Service compatibility is based upon policy: Information such as service-level agreements and security policy isn't captured by a service's WSDL. Instead, tools such as WS-Policy are used to capture information on service policies.

### 4. BUSINESS ACTIVITY MONITORING FRAMEWORK

In this paper, we proposed a framework based real-time data warehouse architecture, as shows in Figure 2.

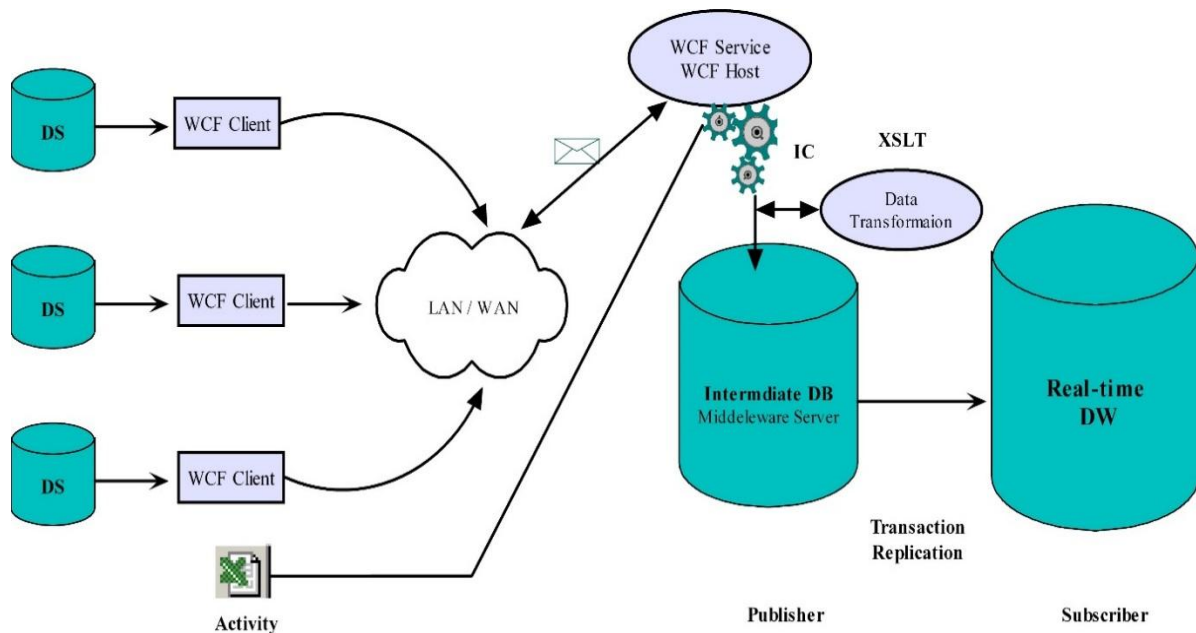


Fig 2: Architecture for real-time data warehousing

Our architecture shows the Phases of real-time as follow:

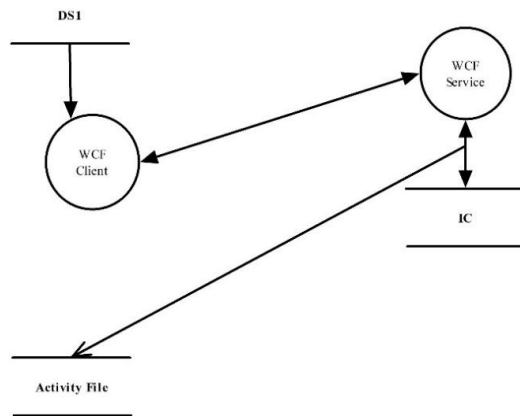
- 1) Data refining  
Data is crude and disorganized. It lacks the structure required for building models or performing analysis. The data refining includes extracting features for unstructured text, combining diverse data sources, filtering interest, selecting relevant features and outcomes for modeling. In this phase an activity is created that define the data to be captured from data sources and how it will be displayed.
- 2) Real-time data tracking  
In this phase the WCF service is created for each data source and service hosting, whereas define the important functions that serve the real-time business need such as data inserted, updated and deleted. Then create an interceptor configuration file (IC) that defines where the data will be captured from. IC is an XML file that specifies the elements for the WCF service should be captured in the activity. Create WCF-client at each data source application. After implement the activity and IC, the framework can track the WCF service, then applying the next phase on tracking inserted, updated and deleted data
- 3) Data Transformation  
Proposed framework need an intermediate DB as staging area before aggregating data and load it to the data warehouse. In this phase applying data transformation process that consists of multiple steps where each step may perform schema and instance related transformation (mappings) [9,10] according to data in the activity file. In metadata and data warehouse, a data transformation converts a set of data values from the data format of a source schema into the data format of a target schema. Data cleaning, also called data purging or scrubbing, deals with detecting and removing blunders and inconsistencies from data in order to enhance the quality of data. Defining the correspondences

between XML schema for different data sources and the intermediate DB table's schemas in a mapping rule definition file. The mapping process designed as XSLT (Extensible Stylesheet Language Transformations) for all transformation tasks. Then the tracking data are inserted into the intermediate DB whereas the intermediate database is counterpart to the real-time DW to eliminate a single point of failure.

- 4) Transaction replication (real-time data load)  
After the data is moved to the intermediate DB. Needing the incremental changes to be propagated from intermediate DB to Real-time DW . Then applying the transaction replication process between the intermediate DB as publisher and the Real-time DW as subscriber.

## 5. CASE STUDY

In this section, an example business process for a multinational manufacturing company called "adventureworks" is presented. The data stores implemented in two servers; DS1, DS2 and DS3 in server 1 . Implementing the intermediate database in middleware server 2 and the real-time data warehouse in server 3. Building the WCF service and its hosting in server2 and the WCF client at each data store applications. These servers are connected to each other by the web or by a high-bandwidth private network, with a communication infrastructure that enables them to conduct Business-to-Business (B2B) transactions. Now need to monitoring the changed data in data stores, in preparation for transfer it to real-time data warehouse. This is done, by performing some steps as shown in Figure 3.



**Fig.3 : Steps of monitoring changed data using WCF**

Creating an activity file for each data store that define and determine the data to be captured and creating the IC file that catches data which has already been identified in activity file going from data store application and show it's in BAM portal. After catching data , applying a real-time data cleaning, integration and transformation by using schema matching. And move this data to the intermediate DB. Hence it is automatically transferred to the real-time data warehouse using transaction replication.

## 6. CONCLUSION

A process data store is fully integrated with the operational environment for business processes and is a data basis that supports these processes with a real-time business intelligence and data warehouse. Proposed framework system design procedure that include the all parts of a real-time ETL as fully integrated. In this paper, the requested data is identified which is refreshed in the real-time. Our framework monitors the changed data using a BAM that provides a real-time business intelligence by capturing this data as it flows through a business system by WCF service. The provided framework designed for minimal impact on data source systems because the needed data is only extracted from sources frequently.

## 7. REFERENCES

- [1] Patricia Ward and George Dafoulas, Database Management Systems.: Thomson Learning, 2006.
- [2] William H. Inmon, Building the Data Warehouse.: Wiley, 2005.
- [3] M. J. Eccles, D. J. Evans and A. J. Beaumont, "True Real-Time Change Data Capture with Web Service Database Encapsulation," 2010 6th World Congress on Services, Miami, FL, 2010.
- [4] B. Wetzstein, P. Leitner, F. Rosenberg, I. Brandic, S. Dustdar, and F. Leymann, "Monitoring and Analyzing

Influential Factors of Business Process Performance," in 13th IEEE Int. Conf. on Enterprise Distributed Object Computing Conference EDOC'09). IEEE, 2009.

- [5] C. Costello, O. Molloy, K. Hinkelmann, "Towards a Semantic Framework for Business Activity Monitoring and Management", Papers from the AAAI Spring Symposium AI Meets Business Rules and Process Management, 2008.
- [6] M. zur Muehlen, R. Shapiro, J. vom Brocke, M. Rosemann, "Business Process Analytics" in Handbook on Business Process Management, Berlin:Springer, vol. 2, pp. 137-157, 2010.
- [7] Janina Voigt. Improving Object Oriented Software Contracts. a thesis submitted in Computer Science. May 2011.
- [8] Ning Wang, Zhiyi Yang and Yalei Yang, "Based on event-driven and service-oriented architecture business activity monitoring design and implementation," 2011 International Conference on System science, Engineering design and Manufacturing informatization, Guiyang, 2011.
- [9] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4):3-13,2000.
- [10] R. Kimball and J. Caserta, The Data Warehouse ETL Toolkit:Practical Techniques for Extracting, Cleanin. John Wiley & Sons, 2004.
- [11] Obali, M., Dursun, B., Erdem, Z., Grr, A.K.: A real time data warehouse approach for data processing. In: Signal Processing and Communications Applications Conference SIU 2013 21st, pp. 1-4 , 2013.
- [12] Lebdaoui, I., Orhanou, G., ElHajji, S.: Data integrity in real-time data warehousing. In: Proceedings of the World Congress on Engineering, 2013.
- [13] Ferreira, N., Martins, P., Furtado, P.: Near real-time with traditional data warehouse architectures: factors and how-to. In: Proceedings of the 17th International Database Engineering & Applications Symposium, IDEAS'13, New York, NY, USA, pp. 68-75. ACM 2013.
- [14] Lebdaoui, I., Orhanou, G., Elhajji, S.: An integration adaptation for real-time datawarehousing. Int. J. Softw. Eng. Appl. 8(11), 115-128 , 2014.
- [15] Wibowo, A.: Problems and available solutions on the stage of extract, transform, and loading in near real-time data warehousing. In: 2015 International Seminar on Intelligent Technology and Its Applications (ISITIA), pp. 345-350, 2015