A process-centric data analytics architecture

Article · February 2015

DOI: 10.1109/INFOS.2014.7036705

CITATIONS
3

READS
182

3 authors:

Mohamed Zaghloul
Mansoura University
6 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Amr Ali-Eldin
89 PUBLICATIONS **740** CITATIONS
SEE PROFILE

Mofreh Salem
Mansoura University
33 PUBLICATIONS **166** CITATIONS
SEE PROFILE

A Process-Centric Data Analytics Architecture

Mohamed M. Zaghloul, Amr Ali-Eldin, and Mofreh Salem
Computer and Systems Department,
Faculty of Engineering,
Mansoura-Egypt
Mohamed.zghloul@huawei.com, Amr.thabet@mans.edu.eg, Moferh.salem@mans.edu.eg

***Abstract*—in this paper, we will present a new approach for Data Analytics development. This new approach is based on the consolidation of the organizational procedures data following a process-centric approach. Three main components are involved in this approach: ETL Component, Process Flow Component and a Control Model Component. Furthermore, we explain the architecture of the process-centric framework gathering these components. Some components of the proposed framework are already deployed in different sites, which will be discussed in detail in the paper. The obtained results showed an obvious enhancement in the CDRs processing and operation cost compared to the traditional files processing approach.**

Keywords— Extraction, Transformation, and Loading (ETL), data analytics , Process-centric collaboration, Self-service data analytics, Control model, Operational Data Store (ODS), data analytics

I. INTRODUCTION

Enterprises have reached a tipping point as their demand for analytics is being driven by its business value as opposed to its technical value. In the early days of data warehousing, technology ruled. Now, business value is the priority. The tectonic shift of business, driving the use of analytics, has been accompanied by two trends that enterprises need to address; data is changing and expanding (In variety, volume and velocity); Analytics uses and users have split into two camps. The potential business value of analytics has expanded across different industries, business functions and enterprise of all sizes. As analytics grow in business value, more people use it – and do so in new ways. Data analytics was once the sole province of business "power" users, but the groups of business people using analytics has bifurcated into different camps: typical analytics and advanced analytics [1], [2].

The typical analytics enable business people to be more self-sufficient because they provide the freedom to explore business data in more ways without needing to request an IT person to create yet another variation of an existing production report. [3]. The advanced analytics camp faces a data integration shortfall, which needs IT support [3].

This paper is trying to propose a new self-service analytical environment for the IT and business users to meet the changing analytics requirements and to provide easy administration, data blending, and data visualization. The proposed environment seeks to empower IT users in their data management operations, and to empower business people to become more self-reliant and less dependent on the IT organization.

This paper is composed of the following sections: Section I presents a short introduction about the paper purpose. Section II presents a literature review. Section III discusses the research challenges, and describes the traditional architecture challenges. The proposed architecture is presented in section IV and V. The experimental work of the proposed framework is described in Section VI followed by an analysis of the results and a discussion in section VII. Conclusion and future work are finally summed up in section VIII. References are listed in section IX.

II. RELATED WORK

The most important design approaches used to develop data analytics solutions are: the traditional approach "bottom-up" ("data-centric") approach and "top-down" ("user-centric") approach. The top-down, report-driven environments require developers to know in advance what kinds of questions casual users want to ask and which metrics they want to monitor. In bottom-up approach developers create a data warehouse model, build extract, transform and load (ETL) routines to move data from source systems to the data warehouse[4], [5], [6].

To succeed in the next decade, data analytics professionals need to adopt a new thinking methodology and new approaches. They need to break away from the "one size fits all" architecture of the past. To meet emerging business demands, they need to manage multiple domains of intelligence and their associated architectures. Without a flexible approach of data architecture, data analytics professionals will overrun with requests and will become the victims of incessant "end-around" plays in which business analysts and departments build their own reporting and analysis environments without the blessing or support of the corporate data analytics team.

III. LIMITATIONS AND CHALLENGES

Therefore, what are the challenges that meet business users while using the traditional data analytics systems?

- Changing Business Demands: data analytics is accused of not being able to cope with the business analytics requirements that continuously change. [3].
- High Operational Costs: Evolving business changes and new analytics needs also lead to issues related to ETL operation costs (the technique responsible for extracting data and transforming it into information) [3], [7].
- Multiple Software Applications: To produce the different analytics needed by the business, corporates use different data analytics software applications that are not gathered in one framework.
- Advanced Analytics Processing Difficulty: data analytics systems require supporting more complex and extreme analytics workloads, which drives up costs [3], [7].
- Big Data Challenges: a tremendous increase in the volumes of data (big data). This data is used for complex, advanced, embedded and streaming analytics. [8], [9], [10].

In this paper, we will consider the detailed explanation of how the proposed framework will meet challenges 2 and 3, the other challenges will be discussed in another paper.

IV. Process-centric Methodology

The idea of the process-centric approach is to gather all the required data analytics software under only one process called "Main Process Flow" to deliver the needed analytics. The main process flow contains sub process flows, where each sub process has a role within the main flow, or acts as an event that feeds external systems. The main process flow handles the integration of all activities as well as the interaction between the different events by developing a metadata repository for all components used within the main process flow. Metadata is related to each activity and event that will be triggered whenever the execution takes place. A Control Model orchestrates all these activities and events in the main process flow. The process-centric approach is designed to work out of technology dependency to spare customers, who already have deployed data analytics software, the hassle of purchasing a specific type of data analytics stack to deploy the process-centric framework.

In order to achieve data transformation, we propose a new customizable ETL (Map) that offers a programmable and integrated sub process for managing data that is now considered either activities or events within the main process flow.

In order to achieve the visualization capability for IT and business users, two intuitive graphical applications are developed to present the front-end of the suggested framework. The first application targets business users who need to develop their own analytics, while the second targets IT users who are interested in executing and monitoring the technical functions of the proposed framework in an easy and robust approach. The provided applications allow information workers and decision makers the mobility they need to consume data wherever they are, and enable IT users to manage the complex operations of data management (Data integration, data transformation, etc.) in an easy manner.

V. Proposed Process-centric Architecture

The proposed process-centric framework contains four processes embedded within the main process flow as shown in Fig. 1.

- The Data Management process handles the data transformation function.
- The Information Delivery process handles the analytics function.
- The Information Insight process handles the detection of hidden patterns in information.
- The Actionable Information process handles providing recommendations to business users.

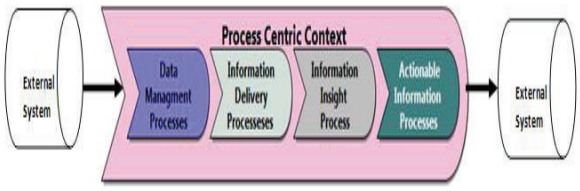

Fig. 1. Process-centric Context

The conceptual process-centric functions show how a process-centric framework will be developed using metadata. The processes' modeling and execution must be implemented in accordance with the specified data model. In particular, the main processes' functions are

- Definition Process: Responsible for identifying the metadata attributes used for defining a given process
- Integration Process: Responsible for identifying the metadata of the process used for integration with the other sub-processes
- Execution Process: Responsible for Identifying the metadata attributes used for executing the integration of the sub processes within the main flow
- Monitoring Process: Responsible for using the metadata attributes to monitor the processes' execution status as shown in Fig.2 Process Centric Function

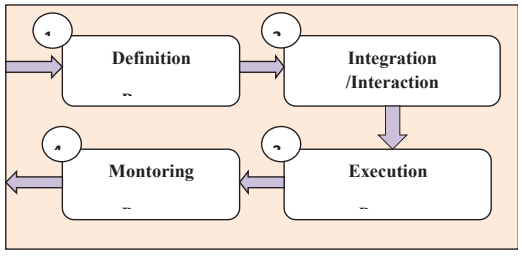

Fig. 2. Process Centric Function

The processes' related activities or events are treated as sub processes within the main process flow. These sub processes are managed and controlled by a Control Model that uses the metadata repository. Not all activities or events should include a customizable ETL as this depends on the purpose of this activity or event. Customizable ETL is used only when data management (data transformation) functions are required.

The Control Model is responsible for managing and monitoring all sub processes levels (activities, events) in addition to the main process flow. According to Fig.3, all the framework's functions are integrated within one process flow as mentioned before. The Control Model enables the management and monitoring of all the activities executed by the framework in order to eventually generate the analytics required by the business users.

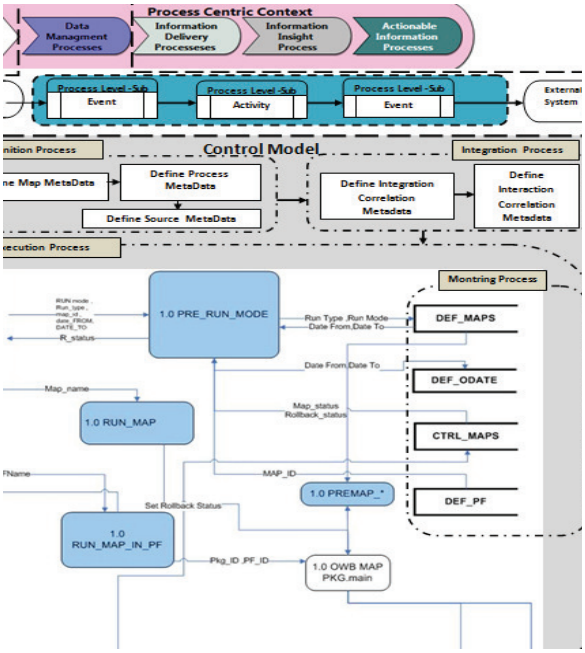

Fig. 3. Process-centric Control Component

The Control Model consists of a number of procedures in addition to a metadata repository including all the sub processes' items, functions, logs, as well as the integration activity's events and rules. The Control Model is optimized for the proposed framework to acquire data resulting from the interaction of the aforementioned resources functioning within the system. After adding the Control Model, the system automatically judges the jobs allocation, and whether a job needs a multi-node parallel processing or not. If necessary, the system divides the data stream into small streams, and then executes the process.

The functions assumed by the Control Model can be summarized in the following points:

- Managing and monitoring the flow of each process.
- Providing the metadata required for each process flow.
- Scheduling and monitoring the process flow in order to eliminate errors that might occur when running different software tools within the same flow.
- Rollback data easily as the monitoring is already executed on the record level.
- PRE_RUN_MODE Procedure Data Flow: the Pre_Run_Mode procedure is responsible for setting the map configuration (setup) data.
- RUN_MAP_IN_PF Procedure Data Flow: the Run_Map_IN-PF procedure is responsible for running a map in the process flow and for storing the run information in the Run Monitoring database tables.
- RUN_MAP Procedure Data Flow: the Run_Map procedure is responsible for running the map and storing the run information in the Run Monitoring database tables.
- PREMAP_* Procedure Data Flow: the PREMAP_* procedure is responsible for storing map run monitoring information, generating new map running sequence, then inserting this sequence (record) into the CTRL_MAPS table, for example: PREMAP_Subscriber.

Another part of the framework is the application targeting IT and business users. The main purpose of such application is to provide a single point of access that allows IT or business users to monitor, run, control and configure most of the proposed framework functions.

Using such application makes the daily operational tasks hassle-free activity and leverages the data warehouse capabilities, introducing a smoothly running process with low operations cost. It also ensures the customers' satisfaction towards their operational experience. Through the IT application, they can:

- Access one interface through which all mappings, process flow components and metadata definitions can be managed and monitored,
- Define map components and process flows settings,
- Have a central point for metadata control and monitoring,
- Through the ETL Scheduler, one of the application functional capabilities, users will be able to:
    - Create new jobs (scheduled tasks) for process flows, dimensions, fact tables and aggregations
    - Set the new ETL jobs to run either manually or automatically according to different time frequencies: by seconds, by minutes, by hours, by days, by weeks, by months and by years.
    - Roll back ETL processes (aggregations and fact tables only).

VI. EXPERIMENTAL WORK

The suggested framework is deployed at some telecom operators that need to process large amounts of CDR (Call Detail records generated from network switches) files. In our case, we have eight CDR types that generate each 50,000 files per day, and we need to process all these types of CDRs, a total that makes around 400,000 files.

A. Experimental Work Objectives

The objectives of using the process-centric approach in CDR Processing as a source for the data warehouse model:

- Empower the IT users by enabling more control capabilities. The control model allows monitoring the CDR processing operations to be used as a source of information for loading the data warehouse model.
- Enhance the data manipulation performance using the process-centric framework and the control model that process more than 400,000 files .

B. Customer Pains

The process complexity is represented by the huge number of CDR files to process and the Java/C programming that needs to be tuned to handle changes in these files. This leads to the necessity of developing a multi-threading methodology capable of dealing with this amount of files not to mention the complexity of tuning their processing in the memory. Fig.4 shows the files' processing flow and the application used to monitor and execute this process.

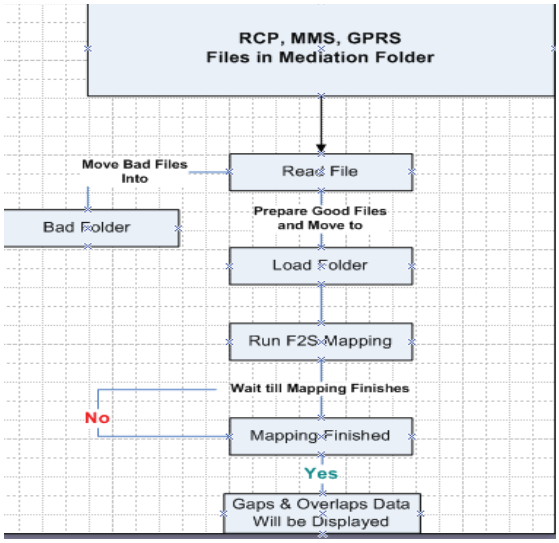

Fig. 4. CDR Processing Flow Chart

Fig.5 shows the traditional application used in the CDR processing. This application will feed the staging tables that represent a source for the data warehouse model.

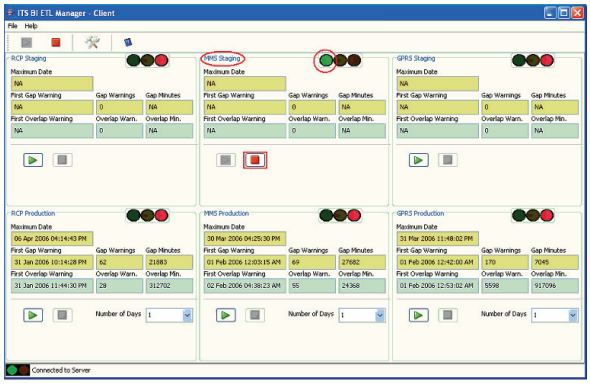

Fig. 5. Traditional CDR Processing Application

C. Experimental Work Specifications

1) Software Specifications:

- Database: Oracle DB 11g Rel. 2
- ETL Tool: Oracle Warehouse Builder 11g Rel. 2.0
- Programming Language: Java Programming
- Java Virtual Machine JDK 1.5.

2) Hardware Specifications:

- Data warehouse Server Processors: 32 processors
- Data warehouse Server RAM:96 GB
- Data warehouse Server Storage: 6TB
- BI Server Processors: 32 processors
- BI Server RAM: 64 GB
- BI Server Storage: 4TB

D. Proposed Architecture

Fig.6 shows the process-centric framework proposed for the CDRs processing. The Files loader is used to integrate the ASCII Files, structured ASCII, binary Files, CDRs, etc… The loader feeds the DB Loader, mentioned hereinafter, with different sources' types. The Files Loader processing is the technique used to transform data that exist in the flat files into database tables to be processed by ITS RA solution later on.

1) Files Loader Features

- Bulk Files Processing
- Keep PLSQL errors Logs into Ctrl's/Def tables.
- Keep original files in one of the ARCH or BAD location as per the result of the loading process.
- Keep BAD files generated from the BD external table engine.
- Keep Log files that show logs for each bulk of files under processing.
- Using standard daily subfolders for Source, Arch, Bad, Logs.
- Keep daily loading in a configuration table

- Keep history of the number of processed files for each day while registering the time and count of files types.
- Ability to process files in parallel which saves a considerable amount of time.
- Ability to Run on UNIX and Windows Operating systems.

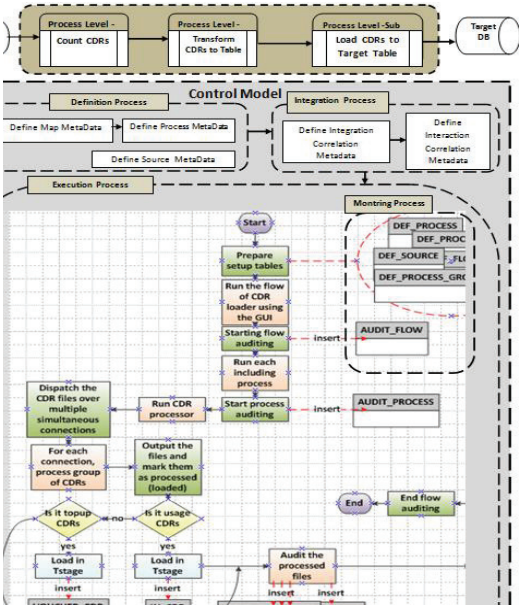

Fig.6. CDR Process-centric Approach

2) CDR Loader Architecture

The File loader as shown in Fig.7 is used to integrate the ASCII Files, structured ASCII, binary Files, CDRs, etc.… The loader feeds the Target tables , mentioned hereinafter, from different sources. The Files Loader processing is the technique used to transform data that exist in the flat files into database tables to be processed using the proposed process-centric approach.

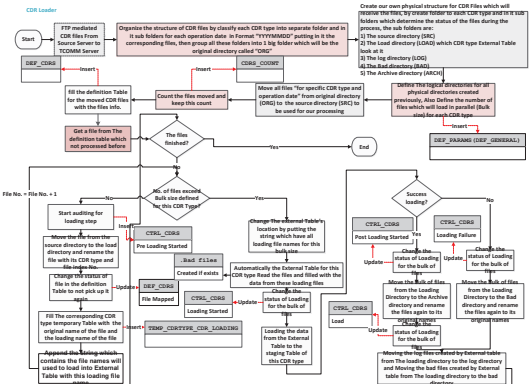

Fig.7. CDR Loader Architecture

E. Experimental Work Results

Some results relative to the deployment of the proposed framework are illustrated herein.

1) Daily Loading Performance

The results come from the deployment of the proposed framework in different customers' sites. Fig.8 illustrates the daily CDR loading performance using the proposed process-centric approach versus the traditional framework.

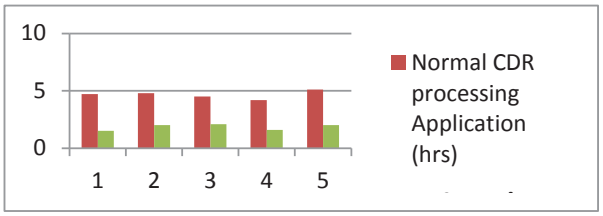

Fig.8. CDR Loading Response Time Using the Proposed Framework

2) Operational Cost Trends

This section illustrates the time needed to support the traditional architecture per month. In the proposed framework, the control model enables the IT to discover the operational errors in a fast and easy manner as errors are tracked on the record level as previously mentioned.

So in brief, IT users have full logs about each record; from where the record is inserted, from which map, from which process flow, at what time, etc. The Control Model allows IT administrators to roll back the wrong set of records and re-load this set if needed. Fig.9 illustrates the operations time consumed when applying the proposed process-centric framework vs. the traditional framework.

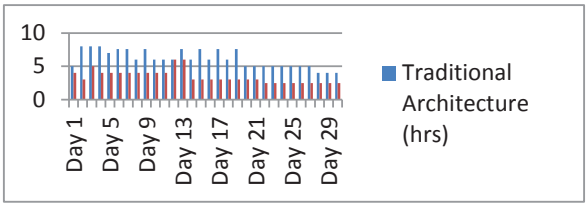

Fig.9. Operation Time Trends

VII. Results Analysis and Discussions

Below is the analysis of the experimental work results:

- Overcoming operations' complexities included in the normal CDR processing approach.
- Data loading errors are easily discovered, and there is no need anymore to load the whole data set again.
- Providing a visual interface for IT users to manage the operational activities. This application can be deployed on a mobile platform and can be configured to send notifications with the loading status.

VIII. Conclusion and Future Work

In this paper, we proposed the Self-service data analytics framework based on a model driven process-centric collaboration architecture. Furthermore, we argued the need for

mainly two components: A model-driven and process-centric Control Component, and a Process Flow Component. We also explained the architecture of these two components. The obtained results show an obvious enhancement on both the performance and operations levels. This is the result of applying the process-centric approach in CDR processing versus the traditional approach. In the future, we will consider handling Big Data within the proposed framework as well as deploying the framework in a cloud-based environment.

IX. REFERENCES

[1] Evelson, B. June 12, 2012. The Forrester Wave™: Self-Service Business Intelligence Platforms, Q2 2012. Forrester.

[2] Business-Software.com. 2012 Edition. Top 10 Business Intelligence Software Report. Business-Software.com.

[3] Imhoff, C. and White, C. Third Quarter 2011. Self-service Business Intelligence: Empowering Users to Generate Insights. TBII Research.

[4] Akkaoui, Z. E., Muñoz, E. Z. J.-N., and Trujillo, J. A Model-Driven Framework for ETL Process Development In Proceedings of the international workshop on Data Warehousing and OLAP. Glasgow, Scotland, UK, pp. 45–52, 2011.

[5] Awad, M. M. I., Abdullah, M. S., and Ali, A. B. M. Extending ETL framework using service oriented architecture. Procedia Computer Science vol. 3, pp. 110–114, 2011.

[6] ISSN 2090-4304 Journal of Basic and Applied Scientific Research www.textroad.com J. Basic. Appl. Sci. Res., 2(1)54-59, 2012 © 2012, TextRoad Publication

[7] Sherman, R. 2012. A Better Way to Fuel Analytical Needs. Sponsored by Composite Software, White Paper.

[8] K. Bhattacharya, R. Hull, and J. Su 2009 . A Data-centric Design Methodology for Business Processes. In J. Cardoso and W.M.P. van der Aalst, editors, Handbook of Research on Business Process Management. Information Science Publishing, an imprint of IGI Global, Hershey, PA, USA.

[9] D. Calvanese, G. De Giacomo, R. Hull, and J. Su 2009. Artifact-centric workflow dominance. In Proc. Intl. Conf. on Service Oriented Computing (ICSOC).

[10] Eckerson, W. 2012. Business-driven BI: Using New Technologies to Foster Self-Service Access to Insights. Tableau Software.

[11] Emerson, M. 2011. Embedding BI Into Your Software Solution — Best Practices. White Paper.

[12] Fields, E.and Sheppard, B. 2012. A New Approach to Business Intelligence: Rapid-fire BI. Tableau Software, White Paper.

