On-Demand ELT Architecture for Right-Time BI:
Extending the Vision

Florian Waas, EMC Corporation, Hopkinton, MA, USA
Robert Wrembel, Faculty of Computing, Poznań University of Technology, Poznań, Poland
Tobias Freudenreich, Department of Computer Science, Technische Universität Darmstadt, Darmstadt, Germany
Maik Thiele, Database Technology Group, Technische Universität Dresden, Dresden, Germany
Christian Koncilia, Institute for Information-Systems, University of Klagenfurt, Klagenfurt, Austria
Pedro Furtado, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal

ABSTRACT

In a typical BI infrastructure, data, extracted from operational data sources, is transformed, cleansed, and loaded into a data warehouse by a periodic ETL process, typically executed on a nightly basis, i.e., a full day’s worth of data is processed and loaded during off-hours. However, it is desirable to have fresher data for business insights at near real-time. To this end, the authors propose to leverage a data warehouse’s capability to directly import raw, unprocessed records and defer the transformation and data cleaning until needed by pending reports. At that time, the database’s own processing mechanisms can be deployed to process the data on-demand. Event-processing capabilities are seamlessly woven into our proposed architecture. Besides outlining an overall architecture, the authors also developed a roadmap for implementing a complete prototype using conventional database technology in the form of hierarchical materialized views.

Keywords: Business Intelligence (BI), Data Stream Analysis, Data Warehouse, Extract-Load-Transform (ELT), Extract-Transform-Load (ETL), Near-Real Time BI, On-Demand ETL, Right-Time BI

DOI: 10.4018/jdwm.2013040102
INTRODUCTION

Business Intelligence (BI) has long been considered an integral part of any successful enterprise’s data processing and analysis strategy (Chaudhuri et al., 2011). BI analysts inspect and query the data, made available through a Data Warehouse to gain insight into sales data or other business facts that will aid them at making business decisions.

Data warehouses are periodically populated or refreshed with data from Operational Data Stores (ODS), e.g., front-end transaction databases. In most businesses, the freshness of the information available in the data warehouse translates directly into more timely business decisions and competitive advantage. Therefore, it is highly desirable to have data available for analysis at real-time or near real-time, i.e., provide data, so no delay is discernable. The degree of delay acceptable depends on the specific application scenario and actual real-time processing in the sense of sub-second delays is generally not needed. This subjective timeliness requirement is sometimes referred to as right-time BI (Davis, 2006).

The biggest hurdle to satisfying right-time BI latency requirements is the data processing needed to make data available in a data warehouse: the data coming from the ODS infrastructure needs to be processed before it is suitable for BI for a variety of reasons. For example, a data warehouse typically consolidates a multitude of different ODS with different schemas and metadata, hence, all incoming data must be normalized. Also, the ODS may contain erroneous or corrupted data that needs to be cleaned and reconciled. This preprocessing is commonly known as Extract-Transform-Load (ETL): data are first extracted from the original data source, then transformed including normalization and cleansing and finally loaded into the data warehouse. For simplicity, we refer to the entire ETL process as loading in the following, unless indicated otherwise. Figure 1 depicts a typical architecture, including various data sources, an ETL layer, and components of the reporting pipeline.

While database technology for data warehousing has seen tremendous performance and scalability enhancements over the past decade in the form of massively parallel database architectures, ETL has improved in scalability and performance to a much lesser degree. As a result, most BI infrastructures are increasingly experiencing an ingest bottleneck: data cannot be furnished to the data warehouse at the necessary pace and freshness. Clearly, in order to provide near real-time or right-time BI this bottleneck needs to be resolved.
Overview of PAKDD Competition 2007
www.igi-global.com/chapter/overview-pakdd-competition-2007/40409?camid=4v1a