Enterprise Data Warehouse Governance Best Practices

Nayem Rahman, Intel Corporation, Hillsboro, OR, USA

ABSTRACT

Maintaining a stable data warehouse becomes quite a challenge if discipline is not applied to code development, code changes, code performance, system resource usage and configuration of integration specification. As the size of the data warehouse increases the value it brings to an organization tends to increase. However these benefits come at a cost of maintaining the applications and running the data warehouse efficiently on a twenty four hours a day and seven days a week basis. Governance is all about bringing discipline and control in the form of guidelines for application developers and IT integration engineers to follow, with a goal that the behavior of a data warehouse application becomes predictable and manageable. In this article the authors have defined and explained a set of data warehouse governance best practices based on their real-world experience and insights drawn from industry and academic papers. Data warehouse governance can also support the development life cycle, maintenance, data architecture, data quality assurance, the Sarbanes-Oxley (SOX) Act requirements and enforce business requirements.

KEYWORDS

Best Practices, Data Quality, Data Warehouse, Data Warehouse Governance, Data Warehouse Maintenance, ETL, Performance Optimization

1. INTRODUCTION

A data warehouse is considered one of the six major IT infrastructures (Weill, et al., 2002; Rahman, 2013a; Rahman et al., 2014) of business organizations. The warehouse expands as the customer base grows as new business requirements are identified. A data warehouse can contain hundreds of applications and thousands of objects. These objects include base tables, global temporary tables, base views, and business views for analytical purposes, stored procedures, macros, join indexes, and triggers.

Data warehouse governance is all about bringing discipline into coding, design and maintenance standards of a Data warehouse. It also helps in making sure that the design and development allows for reusability and code executions are optimal on the target platform. Such optimization (Rahman, 2013b) helps in running multiple applications on the data warehouse with the funded hardware and software capacity of underlying platforms. This reduces the TCO (total cost of ownership) for an IT organization and helps in capacity planning.

In this article we provide a set of data warehouse governance rules including use of automation tools and control criteria, which helps make ETL (Extract-Transform-Load) development (Kimball, 2013) and object migration flawless and the data warehouse environment stable, more efficient, and maintainable. Well executed governance can help an organization achieve the strategic objective of creating and maintaining a state-of-the-art data warehouse to support the ultimate goal of helping business executives make better strategic and tactical decisions based on the information stored in the warehouse (March and Hevner, 2005; Aiken et al. 2011).
2. LITERATURE REVIEW

Data warehouses have the potential to provide business intelligence solutions for companies looking for competitive advantage (Rahman, 2013a). Fortune 1000 companies make strategic and tactical business decisions using the data warehouse as the central repositories of their enterprise data (Wixom & Watson, 2001). In an enterprise data warehouse new projects land over the years and a lot of enhancement and maintenance activities occur as part of day to day operations. All these activities require new objects installation or changing existing objects in the data warehouse. Given these activities how do we ensure that these day to day activities do not make data warehouse environment unstable, cause data quality issues, and impact analytical activities?

Based on real world observations of data warehousing projects implementation and past research findings (Arnott, 2008; Rahman, 2013a; Aiken et al. 2011; Bellatreche & Kerkad, 2015; Rabuzin, 2014; Zolait, 2012) the authors have determined that certain key areas of data warehouse activities need to be governed in a disciplined way. The authors believe that data warehouse objects development, installations, measurement, data quality monitoring, performance monitoring are critical for data warehouse implementation and maintenance. All these are needed to ensure that an organization can develop superior firm-wide IT capability to successfully manage their IT resource and realize agility (Lu and Ramamurthy, 2011; Mithas et al. 2011; Rahman et al., 2011; Roberts and Grover, 2012; Akhter & Rahman, 2015).

Data warehouse implementation has been a research topic for more than a decade. Most previous work on data warehousing focused on design issues (Rahman, 2013a; Rahman, 2014), data maintenance strategies in connection with relational view materialization (Rahman, 2013b) and implementation aspects (Rahman, 2010a; Rahman, 2014). A lot of research work has been done in the field of data warehouse refresh using ETL (Extract-Transform-Load) tools, with different alternative tools being proposed (Simitsis, et al., 2005). Significant amount of research work has also been done to address the issues of data inconsistency and quality (Ballou & Tayi, 1999; Rahman, 2013a). Towards this endeavor our work we focuses on identifying the best practices that can be followed in building and maintaining an enterprise data warehouse. Following the best practices will help in the maintenance, stability, and quality data refreshes for the data warehouse implementation. A consistent and standard enforcement of these best practices are referred to collectively as Data Warehouse Governance.

Data Warehouse Governance has become an important research agenda lately. This is largely due to rapid growth of data in every company’s data warehouse. To the best of our knowledge, there are few publications (Watson, et al., 2004) available on data warehouse governance. They talk about some aspects of governance parameters, for example, the organization structure (Watson, et al., 2004). We have in turn chosen to address issues from the standpoint of ETL development, batch cycles refresh methodologies, data warehouse performance, stability and maintenance of data warehouses. There are other critical parameters influencing governance for a data warehouse implementation such as Corporate, Regulatory and Information Technology levels of governance. In our approach we focus on ETL design (Kimball, 2013) and development standardization, performance optimization of ETL objects as well as batch cycles, controlling of ETL objects so they follow certain standards and run efficiently. We also focus on metadata driven batch cycle refresh.

3. EVOLUTION OF A DATA WAREHOUSE

As time passes data tends to grow in a data warehouse to a very large volumes. Then there comes the significant challenge of maintaining a data warehouse efficiently in terms of space, performance, scalability, and stability (Feinberg and Beyer, 2010; Rahman, 2010b, Rahman, 2013b). Figures 1, 2 and 3 represent how a real-world production Datawarehouse has evolved over a period of time in terms of space occupied by databases/ DW subject areas, active users and database quantities.
Design of Knowledge Based Analytical Model for Organizational Excellence
www.igi-global.com/article/design-of-knowledge-based-analytical-model-for-organizational-excellence/216837?camid=4v1a

A Non-Radial Input Relaxation Model Approach for Evaluating Input Congestion in Data Envelopment Analysis
www.igi-global.com/article/a-non-radial-input-relaxation-model-approach-for-evaluating-input-congestion-in-data-envelopment-analysis/129074?camid=4v1a