Data vault in data warehouse: advantages and disadvantages

Technology, Computer

Applications Implementation Merits and Limitations of Data vault in Data warehouse

Abstract – Business companies face many challenges in exploiting and analyzing data held in diverse sources. Data vault is the latest data warehouse methodology which caters the business needs of flexibility, scalability, agility and large volume of data storage which the prior existing models fail to bestow. Brief data vault architecture, applications of data vault to improve technologies, merits and limitations of data vault are proposed in the review. Data vault 2. 0 a latest methodology which can overcome certain limitations of data vault is also proposed.

Keywords-Data vault; Data vault 2. 0; Data Warehouse

Data warehouse is a subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making. It is a specially prepared repository of data. While Building a data warehouse, aspects such as data modelling, management of business project, risk management, user or organization requirements must be taken into consideration. A For many years, Data warehouse architecture consisted of Inmon or Kimball methodology. Each methodology design has its own pros and cons but are unable to meet the requirements of handling large volume of data processing and re-engineering of data.

Inmon stated that data warehouse is a duplicate of transactional data that is specially structured for the purpose and analysis and querying. It a data driven model where the data is loaded without knowing in prior the user

information. In this model data warehouse and the data marts are separated and have their own storage, scalability and traceability in response to the user requirements. It is time variant, non-volatile, costly and not user-friendly.

Kimball made an innovative approach by making the data warehouse more user friendly by the concept of dimensional modelling. It is composed of facts and dimension tables which provides user the necessary information for decision making. The Kimball data warehouse is consistent of data marts making the initial cost lesser. The Kimball data warehouse is consistent of data marts making the initial cost lesser.

With large amount of data from multiple sources and regular business rules changes, Inmon and Kimball data modelling approaches become less effective. Hence a better evolved model of data vault is created by Dan linstedt.

The Data Vault is a detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business. It is a hybrid approach encompassing the best of breed between 3rd normal form (3NF) and star schema. The design makes the model efficient to store large volumes of data and changes of business rules do not require changes in the data warehouse hence it is cost efficient and user friendly.

The data sources are in 3rd NF and data marts work in star schema. The data vault components are Hubs, Links and Satellites. Hubscontain the unique list of business keys and depicting core concepts of business such as customer,

sales and are vital to identify and track their information. Business keys should have historical uniqueness. Linksare the connections that relates two or more business keys and other links. The hub imposes the links granularity in relation to the link. Satellitecontain the descriptive data that provides context to hubs and links business keys and contains only one parent table. When data changes occur in data warehouse, the descriptive changes are captured in satellites.

Two major technological works is reviewed where data vault is applied to increase the system and business performances:

1. Droid vault– a trusted data vault for android device and a guaranteed platform that provides sensitive data protection from malicious software's for data owners. The model contains two layers of data storage, the green layer where the secure data are processed and the red layers that processes the unsecure data. Droid vault has three components 1. DPM maintains a secure channel for secure data transfer. The sensitive data are encrypted before the data is sent from the droid vault to the android file system. The Bridae module acts an interface. The I/O module secures the user input and display. A unique public/private key is set up for authentication which is one time registered password in the droid vault for secure data transfer to the untrusted android OS users. The design of DV hence provides confidentiality and integrity of sensitive data. The limitation of the model is: a secure environment will provide limited storage hence the data is moved to the untrusted android filesystem. This downside requires additional extra encryption process in the droid vault.

2. Data vaults – database technology for scientific file repository.

Scientific researches is need of efficient technology to explore and manage high volumes of data storage which is rapidly increasing. Hence a data vault technology for storing large volumes of scientific data is constructed. Metadata managed by workflow systems or the file names let researchers search for data. DBMS can approach this issue by processing information at the data storage site, providing malleable query use to analyze and reduce information to TB of data. The limitations of this approach are 1. it is tedious and costly to load the state of art DBMS and DMS will not support specific scientific domain file formats. The solution to this problem in MongoDB data vault. The data vault components are: 1. the data vault wrapper facilitates communication with metadata external file repositories and data access. The virtual data warehouse structure is managed by the data vault cache. The data vault optimizer searches the best query execution plans. The data vault hold the data in its original place & format and parallelly allows transparent metadata and analysis, access of data using query language. The main advantage is the business rules can be applied in advance before the actual loading of data. Hence data vault provides extended functionality and flexibility.

Structural information is separated from descriptive information for reasons of flexibility and avoidance of re-engineering in the case of a change. The Data Vault allows parallel loading of data and suitable for processing large amount of data when compared to the previous available techniques. The data are not processed or filtered. Change of data are never done and

can process large amounts of data.

Flexible, scalable and integrated data model and agile ready and final exploitation of data is not allowed.

Data Vault is only power user accessible. Data in the Data Vault is not " cleansed or quality checked" and the benefits of data vault are indirect but very real.

Data backup is necessary for the business and more up-front is required for long term payoff. The data vault architecture may be flexible and scalable but does not guarantee the reports generated are fully correct.

Data Vault model introduces many joins and integration problems also unstructured data cannot be processed. Data must be made into information BEFOREA A A A delivering to the business.

Data vault 1. 0 is devoid of snowflake data model. Most of source data typically are not tracked by change hence forcing a re-load.

Metadata must be defined in column based level for the business for storage paradigm to make sense. As the complexity of business rules increases the possibility of parallelism and scalability decreases.

Scalability decrease will result in more time consumption for amount of information that is passed for processing.

Data vault 2. 0 is the latest data warehousing methodology which is a novel and improved version to overcome certain downside of data vault 1. 0. The advantages are: 1)The must use of hash key as surrogate key enable the flexibility of data loading in parallel due to independent between satellites therefore paving way for usage of unstructured data in data vaults. 2)Data

Vault 2. 0 is zero dependency type architecture. The data across different can be joined easily, hence allowing dataA A A A vault to be built in multiple platforms and can adapt better to changes.

CRITICAL ANALYSIS

Due to demand of processing large volumes of data andA A A A A A A A A A A A A A A continuous changes in the business rules, data vault model is superior to Inmon and Kimball methodologies in terms of flexibility, agility and scalability and cost. The data vault 2. 0 plays a critical role in minimizing certain important drawbacks. Data vault methodology should be more evolved to overcome the current limitations thereby providing better business and user solution.

CONCLUSION

The data vault methodology proves to be an excellent solution for the data warehouse for reasons of agility, flexibility, scalability etc. The data vault design make the model very effective for storing large volumes of data. The technology applications such as droid vault and data vault for scientific repositories have been modelled with use of data vault was benefitted in terms of security and storage and more. The data vault is advantageous but also has its limitations. Some of the important limitations are overcome by the latest data vault2. 0 methodology. The data vault limitations should be overcome effectively by understanding the business and user needs and create more solutions in a cost-effective way in line with requirements.

REFERENCES

Stuart Lewis, Lorraine Beard, Mary McDerby, Robin Taylor, Thomas Higgins, Clarie Knowles, " Developing a Data Vault", The International Journal of Digital Curation, Vol 11, No 1 (2016)

Milena Ivanova, YaAYiz Kargin, Martin Kersten, Stefan Manegold, Ying Zhang, Mihai Datcu, Daniela Espinoza Molina, " Data vaults: a database welcome to scientific file repositories", IEEE, Computing in Science & Engineering (Volume: 15, Issue: 3, May-June 2013).

Lamia Yessad, Aissa Labiod, " Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault", IEEE, System Reliability and Science (ICSRS).

Xiaolei Li, Hong Hu, Guangdong Bai, " DroidVault: A Trusted Data Vault for Android Devices", IEEE, Engineering of Complex Computer Systems (ICECCS), 4-7 Aug. 2014.

V. Jovanovic, D. Subotic, S. Mrdalj, " Data modeling styles in data warehousing", IEEE, Information and Communication Technology, Electronics and Microelectronics (MIPRO)

William Inmon, Derek Strauss, Genia Neushloss, " DW 2. 0: The Architecture for the Next Generation of Data Warehousing".