RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-15

MR. DHIRENDRA
ASSISTANT PROFESSOR
RAMA UNIVERSITY

OUTLINE

- TUNING THE DATA WAREHOUSE
- TESTING THE DATA WAREHOUSE
  - THREE LEVELS OF TESTING
  - DEVELOPING THE TEST PLAN
  - TESTING BACKUP RECOVERY
  - TESTING THE OPERATIONAL ENVIRONMENT
  - TESTING THE DATABASE
  - TESTING THE APPLICATION
- MCQ
- REFERENCES

Tuning the Data Warehouse

- Tuning the data warehouse deal with the measures such as
  - ❑ Average query response times
  - ❑ scan rates
  - ❑ I/O throughput rates
  - ❑ Time used per query (fixed or ad-hoc)
    - ➢ No. of users in the group Whether they use adhoc queries frequently or occasionally at unknown intervals or at regular or predictable times
    - ➢ The average / maximum size of query they tend to run
    - ➢ The peak time of daily usage
    - ➢ The more unpredictable the load, the larger the queries, or the greater the number of users the bigger the tuning task.

- Memory usage per process

Testing the data warehouse

- **Three levels of testing**
  - ❑ Unit testing : each development unit is tested on its own
  - ❑ Integration testing : the separate development units that make up a component of DWH application are tested to ensure that they work together.
  - ❑ System Testing :the whole DWH application is tested together. The components are tested to ensure that they work properly together, that they don't cause system bottlenecks.
- **Developing the Test Plan**
  - ❑ Test Schedule – metrics for estimating the amount of time required for testing
  - ❑ Data Load
    - ➢ How will the data be generated ?
    - ➢ Where will the data be generated ?
    - ➢ How will the generated data be loaded ?
    - ➢ Will the data be correctly skewed ?

Testing the data warehouse

Testing Backup Recovery: To test the recovery of a lost data file, a data file should actually be deleted and recovered from backup. Check that the backup database is working correctly, and actually tracks what has been backed up and where it has been backed up to. If that works then check that the information can be retrieved. Check that all the backup hardware is working: tapes, tape drives, controllers etc. Each of the scenarios indicated below needs to be centered for

1. Instance Failure
2. Media failure
3. Loss or damage of table space or data file
4. Loss or damage of a table
5. Loss or damage of control file
6. Failure during data movement

Testing the Operational Environment : Testing of the DWH operational environment is another key set o tests that will have to be performed. There are following aspects that need to be tested :

Security – document what is not allowed, disallowed operations and devising a test for each

Disk configuration – test thoroughly to identify any potential I/O bottlenecks.

Testing the data warehouse

Scheduler – Given the possibility for many of the processes in the DHW to swamp the system resources if allowed to run at the wrong time, scheduling control of these processes is essential to the success of the DWH.

Management Tools – (event / system / configuration / backup recovery / database)

Database Management

Testing the database : It can be broken down into three separate sets of tests:

- ❑ Testing the database manager and monitoring tools (creation, running & management of the test database)
- ❑ Testing database features (querying / create index / data load in parallel)
- ❑ Testing database performance (test queries with different aggregations, index strategies , degree of parallel, different-sized data sets)

Testing the application

Logistic of the Test (DWH application code, day-to-day operational procedures, backup recovery strategy, query performance, management & monitoring tools, scheduling software)

Multiple Choice Question

1. _____ is a process of determining the preference of customer's majority.
a) Association
b) Preferencing
c) Segmentation
d) Classification.

2.. Strategic value of data mining is _____.
a) cost-sensitive
b) work-sensitive.
c) time-sensitive.
d) technical-sensitive.

3. _____ proposed the approach for data integration issues.
a) Ralph Campbell.
b) Ralph Kimball.
c) John Raphlin
d) James Gosling

4. The terms equality and roll up are associated with _____.
a) OLAP
b) visualization
c) data mart.
d) decision tree

5. Exceptional reporting in data warehousing is otherwise called as _____
a) exception.
b) alerts
c) errors
d) bugs

REFERENCES

- https://www.tutorialspoint.com/dwh/dwh_overview.htm
- http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf DATA MINING BOOK WRITTEN BY Micheline Kamber
- https://www.javatpoint.com/three-tier-data-warehouse-architecture
- M.H. Dunham, " Data Mining: Introductory & Advanced Topics" Pearson Education
- Jiawei Han, Micheline Kamber, " Data Mining Concepts & Techniques" Elsevier
- Sam Anahory, Denniss Murray," data warehousing in the Real World: A Practical Guide for Building Decision Support Systems, " Pearson Education
- Mallach," Data Warehousing System", TMH
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97 S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96 D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.