Student Code: SE63348

Student Name: Nguyễn Hồng Phát

Class Name: SE1262

|  |
| --- |
| Chapter 8  1/ **Question 1**: Match the columns  1. operational infrastructure A. shared-nothing architecture  2. preemptive multitasking B. provides high concurrency  3. shared disk C. single memory address space  4. MPP D. operating system feature  5. SMP E. vertical parallelism  6. interquery parallelization F. people, procedures, training  7. intraquery parallelization G. easy administration  8. NUMA H. choice data warehouse platform  9. UNIX-based system I. optimize for data transformation  10. data staging area J. data movement option |
| **Your answer 1**:   |  |  | | --- | --- | |  | (A ... J) | | 1 | F | | 2 | D | | 3 | J | | 4 | A | | 5 | G | | 6 | E | | 7 | B | | 8 | C | | 9 | H | | 10 | I | |

|  |
| --- |
| 2/ **Question 2**: What are the platform options for the staging area? Compare the options and mention the  advantages and disadvantages. |
| **Your answer 2**: The staging area sits between data sources and the warehouse repositories. There are 3 options:   * Source data platform (usually legacy platform)   + Advantages: Data is already available in the legacy data source, reduce initial cost.   + Disadvantages: It will affects bold performances of OLTP and OLAP, and old platforms would be hard to get support when things went wrong. * Data storage platform   + Advantages: It’s an ideal options for staging area * A separate optimal platform   + Advantages:     - It is easy to optimize the platform for transformations and cleanings     - Specialized tools can be installed     - Keep tract of entire data content in the staging area   + Disadvantages:     - Require lots of works |

|  |
| --- |
| Chpater 9  3/ **Question 3**: Why do you think metadata is important in a data warehouse environment? Give a  general explanation in one or two paragraphs. |
| **Your answer 3**:  Metadata is important in a data warehouse environment because it is is the road-map to a data warehouse. It defines the warehouse objects and acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.  Metadata also helps in summarization between lightly detailed data and highly summarized data. It is used for query tools, extraction and cleansing tools, reporting tools, transformation tools,... Metadata plays an important role in loading functions. |

|  |
| --- |
| 4/ **Question 4**: Indicate if true or false  A. The importance of metadata is the same in a data warehouse as it is in an operational system.  B. Metadata is needed by IT for data warehouse administration.  C. Technical metadata is usually less structured than business metadata.  D. Maintaining metadata in a modern data warehouse is just for documentation. |
| **Your answer 4**:   |  |  | | --- | --- | |  | (T/F) | | A | F | | B | T | | C | F | | D | F | |

|  |
| --- |
| Chapter 10  5/ **Question 5**: Why is the entity-relationship modeling technique not suitable for the data warehouse? How is dimensional modeling different? |
| **Your answer 5**: Entity-relationship modeling technique not suitable for the data warehouse because:   * E-R models are mean to remove redundancy in the data model, facilitate retrieval of individual records having certain critical identifiers in order to optimize OLTP performance. * In dimensional modeling, a model of tables and relations is constituted with the purpose of optimizing decision support query performance in relational databases, relative to a measurement or set of measurements of the outcomes of the business process being modeled. * Every dimensional model is composed of a “fact” table and a set of “dimension” tables. Conformed fact and dimension elements are elements that conform to the enterprises centralized metadata database. |

|  |
| --- |
| Chapter 11  6/ **Question 6**: How does a snowflake schema differ from a STAR schema? Name two advantages and  two disadvantages of the snowflake schema. |
| **Your answer 6**: The main difference, when compared with the star schema, snowflake’s data table is more normalized. Snowflake schema have many tables and a lot of foreign keys, make the diagram look like the snowflake. Star schema have a fact table at the center and some dimension tables around the fact table, make it look like a star.  Snowflake schema Advantages:   * No redundancy, so snowflake schemas are easier to maintain and change. * When dimension table is relatively big in size, snowflaking is better as it reduces space.   Snowflake schema Disadvantages:   * More complex queries and hence less easy to understand * More foreign keys, higher number of join and hence longer query execution time (slower) |

|  |
| --- |
| Chapter 12  7/ **Question 7**: When is a full data refresh preferable to an incremental load? Can you think of an example? |
| **Your answer 7**: Full data refresh is preferable when there are many changes to be update. It also prefered to an incremental load because refresh is a much simpler option than update. The refresh option simply involves the periodic replacement of complete data warehouse tables, you don’t have to devise the proper strategy to extract the changes from each data source.  Ex: When there are more than 40% of change, we should refresh instead of incremental load. |

|  |
| --- |
| Chapter 13  8/ **Question 8**: Give examples of four types of data quality problems. |
| **Your answer 8**: 4 types of data quality problems:   * Contradicting Values. * Violation of Business Rules. * Reused Primary Keys. * Non-unique Identifiers. |

|  |
| --- |
| Chapter 15  9/ **Question 9**: What is meant by slice-and-dice? Give an example. |
| **Your answer 9**:  Slice operation: performs a selection on one dimension of the given cube, resulting in a sub cube.  Ex: Retrive grades of all students in all subject at a particular year from a 10 years report  Dice operation: defines a sub cube by performing a selection on two or more dimensions.  Ex: Retrive grades of male students in those subject that have more than 3 certificates at the last 3 years.  Slice and dice refers to a strategy for segmenting, viewing and understanding data in a database. Users slices and dice by cutting a large segment of data into smaller parts, and repeating this process until arriving at the right level of detail for analysis. |
| 10/ **Question 10**: Discuss two reasons why feeding data into the OLAP system directly from the source  operational systems is not recommended. |
| **Your answer 10**: Two reasons why it is not recommended:   * Business needed to build queries that summarized the data and make management reports. Such queries were extremely slow because they usually summarize large amounts of data, while sharing the database engine with every day operations. Which in turn **adversely affected the performance of operational systems.** * Operational systems **data source often have data normalized in snowflake schema. These type of schema requires lots of join to summarize the data and make a report, which will delay strategic planning process of the enterprise.** |