BUSINESS INTELLIGENCE 2015W

Test 1 - A
December 9, 2015

Note: For MC-questions, you need to get all four answers correct to get five points; otherwise you get zero points for that question (so it's all or nothing).

1. Data Warehouse Definition (5 points)

   A Data Warehouse ...

   - ✓ is a copy of transaction data specifically structures for query and analysis
   - - stores data by operational applications rather than by business subjects
   - - contains data that is updated continuously
   - - stores only the current values of various transactional data

2. Landing Area (5 points)

   A Landing Area ...

   - - is directly used by OLTP applications
   - ✓ is a database that stores a single data extract of a subset of one source database
   - - contains data that is explicitly structures for analytical queries
   - ✓ has a schema that corresponds to the schema of the subset of the source database

3. Staging Area (5 points)

   A Staging Area ...

   - - is a database that supports one or more types of business transactions
   - ✓ is a database with a schema that corresponds to that of the data warehouse
   - ✓ is a database that is used to store data extracts from various Landing Areas
   - - is a database that stores a single data extract of a source database

4. Data Warehouse vs. Data Marts (5 points)

   Which of the following statements are correct?

   - - Each Data Warehouse focuses exclusively on a single business process
   - ✓ There is typically only one Data Warehouse, but more than one Data Mart
   - ✓ Data Marts may use R-OLAP, M-OLAP or H-OLAP technologies
   - - Data Marts aim to collect and analyze information about the whole enterprise

5. Information integration approaches (5 points)

   Which of the following statements are correct?

   - ✓ Federation connects data sources individually
   - ✓ A Data Warehouse usually only supports mono-directional data flows
   - - A Virtual Data Warehouse is neccessarily faster (i.e., has lower query latency) than a full materialized Data Warehouse
   - ✓ Federation can result in a large number of connections (wrappers)

6. OLTP vs. OLAP (5 points)

   Which of the statements about OLTP and OLAP are correct?

   - OLAP systems optimize for many short and "small" transactions
   ✓ OLAP aims to turn raw data into strategic information
   - OLAP systems tend to use normalized schemas
   ✓ Without a dedicated Data Warehouse, long-running OLAP reads may block OLTP writes

7. Data Warehouse Development approaches (5 points)

   Which of the following statements about Data Warehouse development approaches are correct?

   - Waterfall development approaches toward building BI systems aim to respond to changing requirements rapidly and flexibly
   ✓ The Inman development model starts by first building a centralized Data Warehouse (aka. "Corporate Information Factory")
   ✓ Kimball's Data Warehouse lifecycle process directly builds one Data Mart per major business process rather than building a single centralized Data Warehouse first
   ✓ Traditional waterfall development processes in BI are often associated with a long time to value and a high risk of failure.

8. Facts vs. Dimensions (5 points)

   ✓ Dimensions can usually be thought of as "nouns"
   - Facts control the scope of aggregation
   ✓ Dimension tables provide the "business context"
   ✓ Dimension tables are usually relatively "wide" (many columns)

9. Star- vs. Snowflake-Schema (5 points)

   ✓ The normalized structures in a Snowflake schema are easier to update and maintain than the completely denormalized structures in a Star schema
   ✓ Star schemas result in (small) savings in storage space [??? in contradiction to DWH2, slide 16]
   ✓ Snowflake schemas are typically less intuitive and more difficult to browse for end users
   - Star schemas result in degraded query performance due to additional joins

10. OLAP Operations (6 points)

    Note: A single solution is correct for each of the following four questions (1 P for each correct answer).

    Which OLAP operation can be implemented in SQL by removing a group by clause along a dimension hierarchy?

    – *Slicing* – *Dicing* – *Drill – down* ✓ *Roll – up*

    Adding filter conditions in one or more dimension(s) is called ...

    ✓ *Slicing* – *Dicing* – *Drill – down* – *Roll – up*

    Going from a coarser level of aggregation to a finer (more detailed) level is called ...

    – *Slicing* – *Dicing* ✓ *Drill – down* – *Roll – up*

In a Data Warehousing context, what is typically called "pivoting" or "cross tabulation" in spreadsheet software is known as ...

– *Slicing* ✓ *Dicing* – *Drill – down* – *Roll – up*

Adding filter conditions in one or more dimension(s) is called ...

✓ *Slicing* – *Dicing* – *Drill – down* – *Roll – up*

Going from a coarser level of aggregation to a finer (more detailed) level is called ...

– *Slicing* – *Dicing* ✓ *Drill – down* – *Roll – up*

Which OLAP operation can be implemented in SQL by removing a group by clause along a dimension hierarchy?

– *Slicing* – *Dicing* – *Drill – down* ✓ *Roll – up*

11. Horizontal Partitioning (5 points)

Which of the following statements regarding Horizontal Partitioning are correct?

- ✓] Horizontal partitioning is based on the idea of splitting a table into disjoint parts with the same schema
- ✓ Horizontal partitions can be distributed to different machines/cores
- Horizontal partitioning is particularly efficient when only a few attributes of a table are accessed
- Horizontal partitioning is based on the idea of pre-computing aggregate query results that can efficiently answer other queries over a star schema

12. Time Dimension (5 points)

Which of the following statements about the time demension in Data Warehouse applications are correct?

- ✓ Historization tracks changes in attribute values, relations and entities across time in order to facilitate analysis
- Rows in dimension tables are typically directly associated with time
- The data stored in operational systems contains historic rather than current values
- ✓ The historization of fact tables is typically straightforward, but the historization of dimension tables typically poses a significant conceptual challenge.

13. ETL (5 points)

Which of the following statements about ETL are correct?

- ✓ Syntactic harmonization uses mapping tables to handle key disharmonies, encoding dishamonies, synonyms and homonyms
- ✓ Both file and DBMS technologies may be used in the iplementation of an ETL process
- The loading component transfers data from the source system into the staging area
- ✓ Data auditing aims at judging the quality of data

14. Big data definition (4 points)

Although there is no generally accepted rigorous definition of the term "big data", there is a widely cited characterization around "4Vs" (Gartner, Forrester etc). Name the 4 Vs that describe bit data (there is no consensus over the "fourth V", just name one of the proposals).

_____

_____

_____

_____

15. Horizontal vs. vertical scaling (5 points)

What is the difference between horizontal (i.e., scaling out) and vertical sacling (i.e., scaling up)?

16. Hadoop (5 points)

Hadoop is good at ...

- ✓ fast ingest of massive amounts of data
- \- multi-step ACID transactions
- ✓ exploratory problems
- ✓ complex processing of polystructured data

17. HDFS (5 points)

- ✓ HDFS sits on top of the native filesystem
- \- HDFS is optimized for random access
- \- HDFS performs best with a very large number of small files
- ✓ Files stored in HDFS are "write once"

18. MapReduce (5 points)

MapReduce is ...

- \- a query processing system
- ✓ is a framework that localizes Map tasks with data at various nodes in a cluster
- ✓ a distributed programming model for parallel data processing
- ✓ a batch job processing framework

19. HBase (5 points)

HBase ...

- ✓ is based on a column-family oriented data model
- ✓ automatically shards data into regions by key range and assigns them to region servers
- \- is a well-suited for relational analytics
- ✓ is good for variable schemas

20. Hive (5 points)

Hive ...

- \- always stores schema and data in the same location
- ✓ supports typed columns
- ✓ translates ad-hoc queries and analyses into MapReduce jobs
- \- provides low-latency and supports real-time queries