BUSINESS INTELLIGENCE 2015W

Test 1 - B
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 collection of data in support of management's decision making process
   - ✓ has to contain historical data to facilitate analyses of the past and forecasts for the future
   - \- is overwritten continuously as new data comes in
   - ✓ stores data by business subjects rather than by operational applications

2. Landing Area (5 points)

   A Landing Area ...

   - ✓ is a database that stores a single data extract of a subset of one source database
   - \- is directly used by OLAP applications to execute analytical queries
   - \- is a database that supports applications to execute one or more types of business transactions
   - ✓ has a schema that corresponds to the schema of the data warehouse

3. Staging Area (5 points)

   A Staging Area ...

   - ✓ is a database that is used to store data extracts from various Landing Areas
   - \- contains data that is explicitly structures for analytical queries
   - \- is directly used by operational applications
   - ✓ is used as a source to upload data to the data warehouse from

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

   Which of the following statements are correct?

   - ✓ Each Data Mart focuses on a particular business process
   - \- Data Marts typically have to cope with much larger data volumes than Data Warehouses
   - ✓ Data Warehouses typically make use of the most detailed data available whereas Data Marts use only aggregated Data
   - ✓ Data Marts may use R-OLAP, M-OLAP or H-OLAP technologies

5. Information integration approaches (5 points)

   Which of the following statements are correct?

   - \- A Data Warehousing approach toward data integration supports multi-directional data flows from and to individual source systems
   - ✓ Source systems communicate with other source systems in a federated information integration architecture

✓ A virtual warehouse (mediator) turns a user query into a sequence of source queries and assembles the results

✓ Sources in a Data Warehouse are translated from their local schema into an integrated global schema and copied to a central DB

6. OLTP vs. OLAP (5 points)

Which of the statements about OLTP and OLAP are correct?

✓ OLTP systems optimize for many short and "small" transactions

- OLTP typically involves large periodic batch inserts

- OLTP systems store historized data

✓ OLAP aims to turn raw data into strategic information

7. Data Warehouse Development approaches (5 points)

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

- Kimball's Data Warehouse lifecycle process starts by first building a centralized Data Warehouse (aka. "Corporate Information Factory")

✓ A traditional waterfall development approach involves phases such as specification, platfomr/software/tools selection, realization and operation

✓ Kimball's development model involves three concurrent tracks

✓ "Agile BI" is a time-boxed, iterative, evolutionary development approach

8. Facts vs. Dimensions (5 points)

- Facts can usually be thought of as "nouns"

✓ Dimensions control the scope of aggregation

✓ Facts can be aggregated

- Dimension tables are usually relatively "long" (many rows)

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

- Compared to Snowflake schemas, the normalized structures in Star schemas are easier to update and maintain

✓ Snowflake schema result in (small) savings in storage space

✓ Star schemas are typically more intuitive and easier 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).

Going from a finer level of aggregation to a more coarse level is called ...

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

Which operation can be implemented in SQL by adding a group by clause along a dimensional hierarchy?

– *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*

Going from a finer level of aggregation to a more coarse level is called ...

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

Which operation can be implemented in SQL by adding a group by clause along a dimensional hierarchy?

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

11. Materialized Views (5 points)

Which of the following statements about Materialized Views (MVs) are correct?

- ✓] A DBMS's query optimizer may speed up query processing by sourcing MVs instead of tables (e.g., by avoiding joins)
- ✓ MVs trade off speed of query processing vs. storage space
- MVs are based on the idea of splitting a table into disjoint parts with the same schemas
- MVs are based on the idea of dividing a table into multiple tables that contain fewer columns

12. Time Dimension (5 points)

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

- ✓ The concept of "slowly changing dimensions" is based on the notion that the frequency of dimensional changes is usually lower than that of fact changes
- The data stored in operational systems contains historic rather than current values
- ✓ Historization tracks changes in attribute values, relations and entities across time in order to facilitate analysis
- Archiving is used when historic data must be retrieved frequently (e.g. to facilitate analysis)

13. ETL (5 points)

Which of the following statements about ETL are correct?

- Only virtual DWHs need an ETL process
- ✓ ETL monitoring approaches include trigger-based, replication-based, log-based, timestamp-based, and snapshoot-based strategies
- The extraction component transfers data from the staging area into the DWH
- ✓ Data scrubbing uses domain knowledge to detect "dirty" 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 scaling (i.e., scaling up)?

16. Hadoop (5 points)

Hadoop is good at ...

- \- interactive analysis of higly structured data
- \- low latency relational analytics
- ✓ schema-on-read style flexible analysis of "data lakes"
- ✓ providing massive scalability of storage and computation

17. HDFS (5 points)

- \- HDFS is an OS file system that runs on a local machine
- ✓ HDFS is conceptually based on Google's GFS
- \- HDFS performs best with a very large number of small files
- ✓ HDFS is optimized for large, streaming reads

18. MapReduce (5 points)

MapReduce is ...

- \- a distributed data store
- ✓ a distributed programming model for parallel data processing
- ✓ a batch job processing framework
- \- a streaming data processing tool

19. HBase (5 points)

HBase ...

- ✓ facilitates parallelized queries over massive data sets
- \- is a graph database
- ✓ is a key-value data store
- \- databases can be accessed directly via SQL

20. Hive (5 points)

Hive ...

- ✓ translates ad-hoc queries and analyses into MapReduce jobs
- \- is a relational database
- \- is designed for on-line transaction processing
- ✓ stores schema information in a relational database (Metastore)