**44-560 Advanced Topics in Database Systems**

**Introduction to Data Warehousing KEY**

1. Consider the dimensional model given below for a grocery store chain.



Assume that

* + there are 500 stores;
  + there are 200,000 products;
  + data will be collected for 5 years (5 \* 365 days);
  + on any given day at any given store, approximately 5,000 different items are sold.

Answer the questions below. Assume the average field size is 10 bytes.

* 1. How many rows will **DateDim** contain?

5 \* 365 = 1,825

* 1. How many rows will **ProductDim** contain?

200,000

* 1. How many rows will **StoreDim** contain?

500

* 1. How many total bytes of storage will the dimension tables require?

DateDim: 1,825 rows \* 4 fields / row \* 10 bytes / field = 73,000 bytes

ProductDim: 200,000 rows \* 3 fields / row \* 10 bytes / field = 6,000,000 bytes

StoreDim: 500 rows \* 5 fields / row \* 10 bytes / field = 25,000 bytes

Total bytes of storage = 73,000 + 6,000,000 + 25,000 = 6,098,000

* 1. How many rows will the fact table contain?

On any given day, at any given store, approximately 5,000 different items are sold, so

500 (number of stores) \* 1,825 (number of days) \* 5,000 (number of different products sold at a particular store on a particular day) = 4,562,500,000 rows

* 1. How many bytes of storage will the fact table contain?

4,562,500,000 rows \* 6 fields / row \* 10 bytes / field =273,750,000,000

1. Suppose we use a coarser grain and maintain data for stores and dates but not individual products. For each store, we lump all product sales together each day. The new model looks like this:



Answer the questions below. Assume the average field size is 10 bytes.

* 1. How many rows will **DateDim** contain?

5 \* 365 = 1,825

* 1. How many rows will **StoreDim** contain?

500

* 1. How many total bytes of storage will the dimension tables require?

DateDim: 1,825 rows \* 4 fields / row \* 10 bytes / field = 73,000 bytes

StoreDim: 500 rows \* 5 fields / row \* 10 bytes / field = 25,000 bytes

Total bytes of storage = 73,000 + 25,000 = 98,000

* 1. How many rows will the fact table contain?

500 (number of stores) \* 1,825 (number of days) = 912,500 rows

* 1. How many bytes of storage will the fact table contain?

912,500 rows \* 5 fields / row \* 10 bytes / field =4,5625,000

1. Suppose we maintain data for dates and products, but not for individual stores. Then our model will look like this:



Answer the questions below, assuming

* + the average field size is 10 bytes, and
  + on a given day, across all stores in the chain, approximately 10,000 different products are sold.
  1. How many rows will **DateDim** contain?

5 \* 365 = 1,825

* 1. How many rows will **ProductDim** contain?

200,000

* 1. How many total bytes of storage will the dimension tables require?

DateDim: 1,825 rows \* 4 fields / row \* 10 bytes / field = 73,000 bytes

ProductDim: 200,000 rows \* 3 fields / row \* 10 bytes / field = 6,000,000 bytes

Total bytes of storage = 73,000 + 6,000,000 = 6,073,000

* 1. How many rows will the fact table contain?

1,825 (number of days) \* 10,000 (number of different products sold on a particular day) = 18,250,000 rows

* 1. How many bytes of storage will the fact table contain?

18,250,000 rows \* 5 fields / row \* 10 bytes / field =912,500,000

1. Construct a star schema for a national bookstore chain. The relevant dimensions and attributes are as follows:

* **BookDim**. Attributes: ISBN, title, publisher, year of publication.
* **StoreDim**. Attributes: store number, store name, street address, city, state, zip, name of manager, phone number.
* **DateDim**. Attributes: day, month, year, quarter, holiday flag, and weekend flag.

Facts to be recorded are number of copies sold, dollar amount, and dollar cost.

The chain carries approximately 25,000 titles and there are 150 stores nationwide. Each store carries an average of 15,000 titles. On any given day, at any given store, approximately 200 different titles are sold. Data will be stored for 5 years (5 \* 365 days).

For the problems below, assume the average field size is 5 bytes.

* 1. Place your star schema in the space below. Remember to use surrogate keys for the primary keys of the dimension tables.



* 1. How many rows will the **BookDim** table contain?

There are 25,000 titles so there will be 25,000 rows.

* 1. Find the size (in bytes) of **BookDim**.

Each row contains 5 attributes. Average size of an attribute is 5 bytes, so the average row size is 25 bytes. 25 bytes / row \* 25,000 rows = 625,000 bytes.

* 1. How many rows will the **StoreDim** table contain?

There are 150 stores, so there will be 150 rows.

* 1. Find the size (in bytes) of **StoreDim**.

There are 10 attributes in **StoreDim**, of average size 5 bytes each, so the average row size is 50 bytes. 50 bytes/row \* 150 rows = 7,500 bytes.

* 1. How many rows will the **DateDim** table contain?

We are storing data for five years – one row per day – so there will be 5 \* 365 = 1,825 rows.

* 1. Find the size (in bytes) of **DateDim**.

There are 7 attributes in each row, of average size 5, so average row size is 35 bytes.

35 bytes/row \* 1,825 rows = 63,875 bytes

* 1. How many rows will the fact table contain when all five years of data has been stored?

There are 1,825 dates and for each date, we store information for each title sold in each store. There are 150 stores and on any given day, each store will sell about 200 different titles. 1,825 days \* 200 titles sold/store \* 150 stores/day = 54,750,000 titles.

* 1. Find the size (in bytes) of the fact table when all five years of data has been stored.

There are six fields in the fact table, averaging 5 bytes per field, for an average row size of 30 bytes. 30 \* 54750000 = 1,642,500,000 bytes

1. Construct a star schema that enables a university with multiple campuses to track enrollments on the various campuses. The relevant dimensions and attributes are:

* **CampusDim**. Attributes: CampusID, campus name, city, state, zip.
* **StudentDim**. Attributes: StudentID, name of student, street address, city, state, zip, majorID (assume a student has only one major), major name.
* **DateDim**. Attributes: semester (fall, spring, or summer), year.

Facts to be recorded are number of credit hours the student is enrolled in on each campus and the dollar amount of revenue generated for a campus by the student’s enrollment.

The university has 5 campuses. Total student enrollment for the five campuses is approximately 100,000 each semester. Each semester (fall, spring, and summer), about 30,000 students transfer or graduate and are replaced by new students. In any given semester, about 20,000 students are enrolled at two different campuses. Data will be stored for 10 years.

Note that when a student transfers or graduates, that student is *not* removed from the database, because we are storing historical data.

For the problems below, assume the average field size is 10 bytes.

* 1. Place your star schema in the space below. Remember to use surrogate keys for the primary keys of your dimension tables.



* 1. How many rows will the **CampusDim** table contain?

There are 5 campuses so there will be **5** rows.

* 1. How many rows will the **StudentDim** table contain at the end of the ten year period?

During the first semester there are 100,000 students so there will be 100,000 rows. Each semester about 30,000 leave (but remain in the database) and are replaced by 30,000 new students. At the end of 10 years (29 additional semesters), 870,000 new students will have been added for a total of **970,000** rows of data in the **StudentDim** table.

* 1. How many rows will the **DateDim** table contain?

We are keeping data for 10 years. For each year we have three semesters, so we will have 10 \* 3 = **30** rows.

* 1. Find the size (in bytes) of **StudentDim**.

There are 10 attributes in **StudentDim**, of average size 10 bytes each, so the average row size is 100 bytes. 100 bytes/row \* 970,000 rows = **97,000,000** bytes.

* 1. How many rows will the fact table contain when all ten years of data has been stored?

There are 30 dates, representing the 30 semesters. For each date, the fact table will contain about 100,000 students since that is the approximate number of students who enroll each semester. 20% enroll on more than one campus, adding another 20,000 rows. 120,000 rows per semester \* 30 semester = **3,600,000** rows.

* 1. Find the size (in bytes) of the fact table when all ten years of data has been stored.

3,600,000 rows \* 5 fields per row \* 10 bytes per field = **180,000,000** bytes