Part1

Question1

(1). No, authors cannot be the primary key. The primary key has to be unique. But for different books' titles, they may have same authors.

(2). Book3 is capable of answering such query.

SQL query:

```
SELECT Titles, SalesRanks, PaperbackPrice, EbookPrice, AudiobooksPrice
From Book3
WHERE SalesRanks >= 1
AND SalesRanks <= 99
ODER BY SalesRanks;
```

Question2

(1). Vertical fragmentation strategy should be used to fragment Book2 table.

Vertical fragmentation

Fragment1:

(id,book_title,authors,publication_year,publication_month,publication_day,edition,isbn13,language,series,pages)

Fragment2:

(id,publisher_name)

(2). Firstly, we should identify the publication_day of this new record. Then, we should determine which fragment that the new record can be inserted based on the predicate of three fragments. Finally, the new record is inserted into corresponding fragment.

Part2

Question3

(1). The Day, Publisher, Language are dimension columns. The Sales is a fact column.

Question4

(1).

The advantages of bitmap:

(1). The bitmap significantly reduces space and I\O.

(2). The bitmap also reduces query processing time.

Publisher and Language are suitable for bitmap index.

(2).

ID	Publisher			
	AAAI Press	Springer International Publishing	Springer London	IEEE Computer Society Press
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	1	0	0	0
6	0	1	0	0
7	0	0	1	0
8	0	0	0	1

ID	Day	Language	
		English	Spanish
1	07/15/1984	1	0
2	05/05/1990	1	0
3	06/04/1995	1	0
4	12/11/2000	1	0
5	04/03/2004	0	1
6	05/01/2008	0	1
7	11/19/2012	0	1
8	08/06/2014	0	1

(3). Firstly, scan the AAAI Press vector and select all records whose AAAI Press = 1.

ID	Publisher			
	AAAI Press	Springer International Publishing	Springer London	IEEE Computer Society Press
1	1	0	0	0
5	1	0	0	0

Secondly, scan the Spanish vector and select all records whose Spanish = 1.

ID	Language	
	English	Spanish
5	0	1
6	0	1
7	0	1
8	0	1

Thirdly, join these selected records from their respectively vectors. Then find the record that satisfied the requirement is the ID = 5 (the fifth record in the original table) and its corresponding Day is 04/03/2004. Therefore, the total sales of “Spanish” books published by “AAAI Press” is 2

Day	Publisher	Language	Sales
04/03/2004	AAAI Press	Spanish	2

Part3

Question5

(1). Global Conceptual Schema

Books (id, title, authors, date, publisher, ISBN13, pages)

The id of global conceptual schema can be found in Book1.id, Book2.id, Book3.ID and Book4.ID.

The title of global conceptual schema can be found in Book1.title, Book2.book_title, Book3.Title and Book4.Title.

The authors of global conceptual schema can be found in Book1.authors, Book2.authors, Book4.Author and derived in Book3.Author1 & Book3.Author2 & Book3.Author3.

The date of global conceptual schema can be found in Book3.Date, Book4.Publication_Date and derived in Book1.pubyear & Book1.pubmonth & Book1.pubyear, Book2.publication_year & Book2.publication_month & Book2.publication_day.

The publisher can be found in Book1.publisher, Book2.publisher_name, Book3.Publisher and Book4.Publisher.

The ISBN13 can be found in Book1.isbn13, Book2.isbn13, Book3.ISBN13 and Book4.ISBN13.

The pages can be found in Book1.pages, Book2.pages, Book3.Pages and Book4.Pages.

(2).

Data type heterogeneity. Date in different schemas may have different format, like 05/11/2007 or May.11th, 2007. In this case, before we do the integration, we need to check the data type and transform the different date type into same type, like convert them to the same form 01/01/2000.

Semantic heterogeneity. Book3.Date may refer to the publication date of books or the date that the book was sold. In this case, we can sample some books from this schema and search their information on the Internet to confirm the meaning of the Date. If the Date is the publication date, then we can integrate data directly. Otherwise, we need to create a column to store the publication date of these book, then doing the integration.

Part4

Question6(The code of this question is in the src\Q6.py file)

(1). There are 37 records in the sample set.

(2). There are 286 fields containing NULL present.

(3). The Empo is 454689.98.

Question7

(1). Jaccard distance is more likely to regard them as similarity. The edit distance only compares the characters of the string and it does not consider of the meaning of the string. Therefore, even if these two strings have same meaning, the edit distance will be large because the author in these two strings have different orders. But the Jaccard distance is less sensitive to the word orders.

(2). (The code for this question is in the src\Q7(2) file)
Precision=0.1912479740680713,
Recall=0.5086206896551724,
F-measure= 0.27797408716136635.