**Part 1.**  
**Question 1**:

SELECT ROW\_NUMBER() OVER (order by num desc), title, p, pages ,num from

(

select title, publisher as p , pages, COUNT(\*) as num from Book1 b1

group by title

UNION ALL

select book\_title as title, publisher\_name as p, pages ,COUNT(\*) as num from Book2 b2

group by book\_title

UNION ALL

select Title as title, Publisher as p, Pages as pages ,COUNT(\*) as num from Book3 b3

group by Title

UNION ALL

select Title as title, Publisher as p , Pages as pages, COUNT(\*) as num from Book4 b4

group by Title

)

limit 100;

**Question 2**:

**2.1**

Mixed Fragmentation should be used to fragment Book2 table.

Predicates Pr = {publisher\_name=“XXX1”,publisher\_name=“XXX2”… publisher\_name=“XXXn”,, publication\_year==“the given year”, publication\_year!=“the given year”}

and schema divided : [id, book\_title, authors, language, pages] and [ publication\_year, publication\_maonth, publication\_day, edition, publisher\_name, isbn13, series]

If only 2 fragments are generated, I will use Vertical fragment,

the schema is divided into :

[id, book\_title, authors, language, pages] and [ publication\_year, publication\_maonth, publication\_day, edition, publisher\_name, isbn13, series]

**2.2**

not set valid , because is:Disjointness is not fulfilled, as Fragement 2 and Fragent 3 have jointness.

min-term predicate steps:

step1:

p1 : pages ≤ 100

p2 : pages ≥ 100 and pages ≤ 800

p3: pages ≥ 600

step2:

Pr = {p1,p2, p3} which is not Disjointness, Minterm predicates:

m1 : (pages ≤ 100)

m2: (pages ≥ 600 )

m3 : NOT(pages ≤ 100) and NOT(pages3 ≥ 600 )=(pages >100 ) and(pages < 600)=(100<page<600)

then switch m2 and m3 places:

m1 : (pages ≤ 100)

m2: (100<page<600)

m3: (pages ≥ 600 )

step3:

then outcome of M = { m1, m2, m3} ,where m1 : (pages ≤ 100) m2: (100<page<600) m3: (pages ≥ 600 )

if we want to insert a new record into Book2:

1. check new insert data’s pages, to know new record should insert to m1 range or m2 range or m3 range
2. in the range the new record should be inserted, check the fragment table’s status and space, whether the new record can/be inserted, if we can , the insert into it , if now , we should do shrink table and make space for the new record

additon: if we are with and above Oracle11g, because we are range fragment, once a new record in insert, we could use “Interval partition” ,Oracle wiill automatically manag partition.

**Part 2**

**Question 3**:

Fact table has the most records.

Because Fact table associated with each fact is a key identifying, for example, which day, which book and which publisher and which language.

**Question 4**

**4.1**

The advantages of using bitmap indexes are：

Significant reduction in space and I/O  
Reduce processing time：

Comparison, join and aggregation operations can be reduced to bit operation ；

Bit operations are very fast

Bitmap indexs is not suitable for high cardinality domains

**4.2**

publisher language

AAAI Press | Springer International Publishing | Springer London |IEEE Computer Society Press English |Spanish

1 0 0 0 1 0

0 1 0 0 1 0

0 0 1 0 1 0

0 0 0 1 1 0

1 0 0 0 0 1

0 1 0 0 0 1

0 0 1 0 0 1

0 0 0 1 0 1

then the bitmap index is :

1 0 0 0 1 0

0 1 0 0 1 0

0 0 1 0 1 0

0 0 0 1 1 0

1 0 0 0 0 1

0 1 0 0 0 1

0 0 1 0 0 1

0 0 0 1 0 1

**4.3**

bitmap of “AAAI Pres” and language “english” is: 100010

In order to find the total sales of “English” books published by “AAAI Press”:

we can filter out **T** from fact tables by bitmap index = 100010

then Select SUM(Sales) from **T**

**Part 3**

**Question 5**

**5.1**

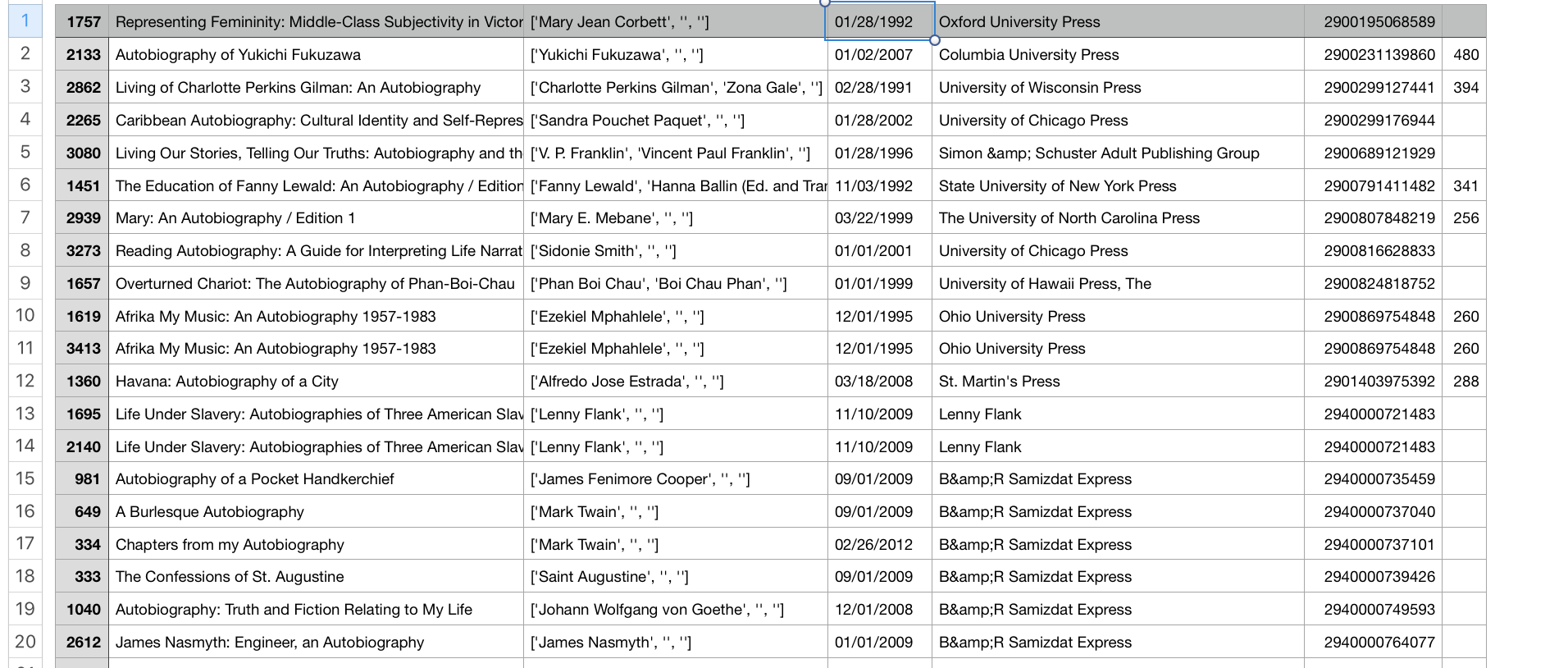
global conceptual schema is:

FullBook(id, title, authors, publish\_date, publisher, isbn13, pages)

**5.2**

code is at: Assignment/src/Q5\_2Integrate\_Book3Book4.py

output csv is at: Assignment/out/FullBookList.csv

****

**Question 6**

code is at: Assignment/src/Q6.py

the linkage-pair file we generated is at: Assignment/out/b1\_b2\_linkage.csv

tp, fp, fn= 212 937 20

precision= 0.18450826805918188

recall= 0.9137931034482759

f\_measure= 0.3070238957277335

**Question 7**

code is at: Assignment/src/Q7.py

field\_count= 62900 error\_count= 28633

error\_count / field\_count= 0.455214626391097