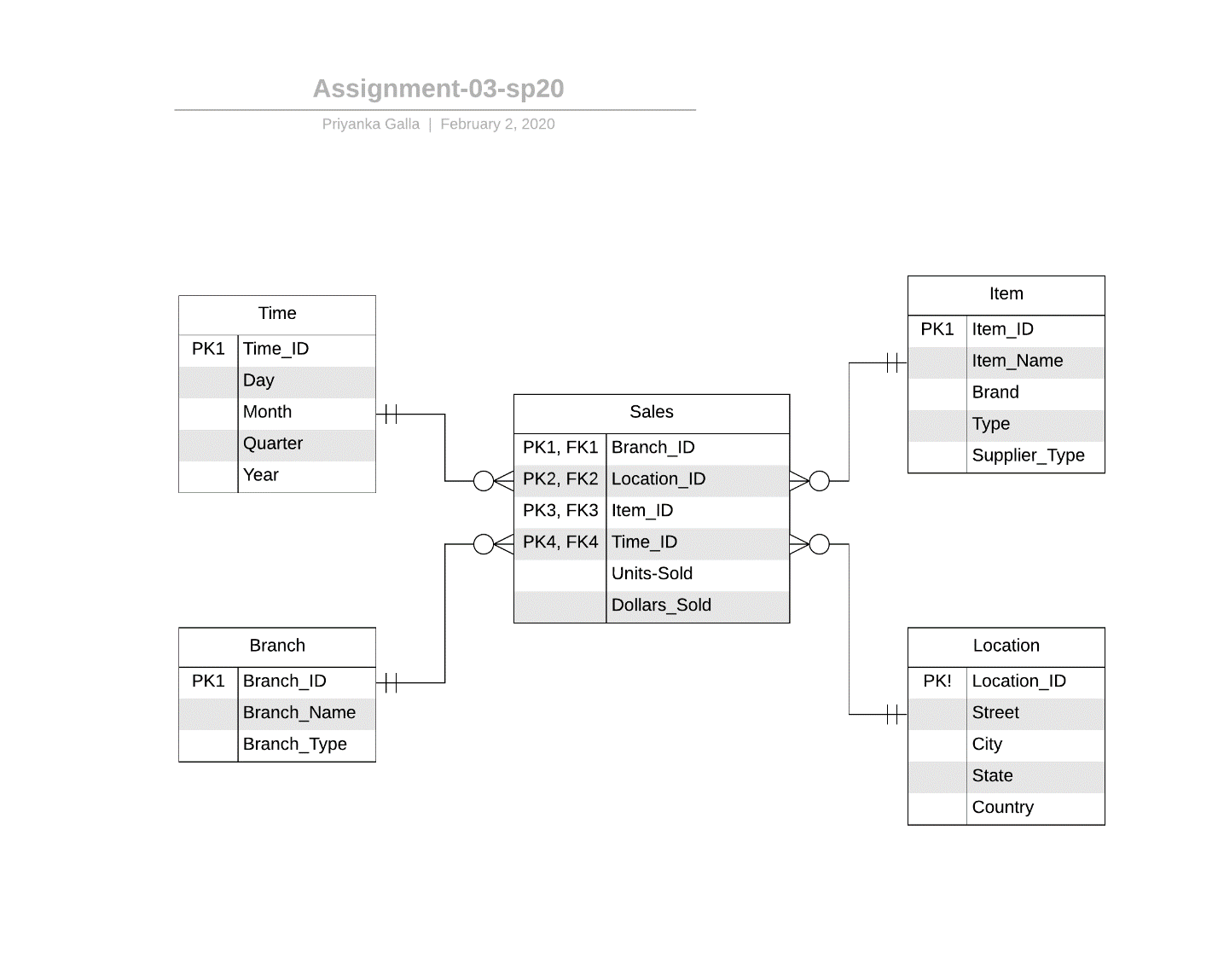
**44-560 Advanced Topics in Database Systems**

**Assignment 05: Database Performance Tuning and Query Optimization**

The problems on this assignment uses the following database:



The above ER diagram is the database design of the Macy’s store. This database models all the sales of different items, on different dates and times at different locations of the store. Assuming the Sales table contains the registered sales at the store, there will be hundreds of thousands of sales occurring in an hour across all the stores in the USA. The data management design must be optimized in order to make the transactions fast and efficient. There are two access plans given below. They vary in the approach of the execution of the same query. You will need to calculate the I/O cost and the number of resulting rows for both access plans and compare them with each other.

Assume that,

* The Item table has 100,000 rows.
* The Location table has 500 rows.
* The Branch table has 60,000 rows.
* The Time table has 35,000 rows.
* The Sales table has 3,000,000 rows.
* The item ‘Hollister women’s cropped tee’ shows up in 7000 rows of the Sales table and only once in the Item table.
* The Time\_ID 040196 shows up in 500 rows of the Sales table.
* The item ‘Puma grey men’s’ shoes’ shows up in 700 rows of the Sales table and only once in the Item table.
* The location ‘Michigan’ shows up in 1000 rows of the Sales table and only once in the Location table.
* The location ‘Chicago’ shows up in 2000 rows of the Sales table and only once in the Location table.

1.) Consider the following query:

SELECT Location.Location\_ID, Location.Country, Location.State, Sales.Sales\_ID, Sales.Units\_Sold, Sales.Dollars\_Earned

FROM Locatoin, Sales

WHERE Location.Location\_ID = Sales.Location\_ID

AND Sales. Time\_ID = 040196;

Below are two access plans for executing this query. Fill in the missing information in the last four columns (assume that the cost of one I/O operation is 1).

Plan A:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Step** | **Operation** | **I/O Operations** | **I/O Cost** | **Resulting Rows** | **Total I/O Cost (accumulated)** |
| A1 | Cartesian product (Location, Sales) | 500 + 3,000,000 | 3,000,500 | 1.5\*(10)9 | 3,000,500 |
| A2 | Select rows from A1 with matching Location\_ID | 1.5\*(10)9 | 1.5\*(10)9 | 3,000,000 | 1,503,000,500 |
| A3 | Select rows from A2 with Time\_ID = 040196 | 3,000,000 | 3,000,000 | 500 | 1,506,000,500 |

Plan B:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Step** | **Operation** | **I/O Operations** | **I/O Cost** | **Resulting Rows** | **Total I/O Cost**  **(accumulated)** |
| B1 | Select rows from Sales with Time\_ID = 040196 | 3,000,000 | 3,000,000 | 500 | 3,000,000 |
| B2 | Cartesian product (B1, Location) | 500 + 500 | 1,000 | 2,50,000 | 3,001,000 |
| B3 | Select rows in B2 with matching Location\_ID | 2,50,000 | 2,50,000 | 500 | 3,251,000 |

2.) Which plan is better to use among the above two plans in question-1? Why?

A. Plan B is better to use because you can clearly see that the “Total I/O Cost” is less when compared to Plan A. We consider the plan which I/O cost is less making it optimize.

3.) Consider the following query:

SELECT Sales.Sales\_ID, Item.Item\_ID, Item. Item\_Name

FROM Sales, Item

WHERE Sales. Item\_ID = Item. Item\_ID

AND Item. Item\_Name = ‘Puma grey men’s’ shoes’;

Below are two access plans for executing this query. Fill in the missing information in the last four columns.

Plan A:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Step** | **Operation** | **I/O Operations** | **I/O Cost** | **Resulting Rows** | **Total I/O Cost**  **(accumulated)** |
| A1 | Cartesian product (Sales, Item) | 1,00,000 + 3,000,000 | 3,100,000 | 0.3\*(10)12 | 3,100,000 |
| A2 | Select rows from A1 with matching Item\_ID | 0.3\*(10)12 | 0.3\*(10)12 | 3,000,000 | 3.00003100(E)11 |
| A3 | Select rows from A2 with Item\_Name = ‘Puma grey men’s’ shoes’; | 3,000,000 | 3,000,000 | 700 | 3.00006100(E)11 |

Plan B:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Step** | **Operation** | **I/O Operations** | **I/O Cost** | **Resulting Rows** | **Total I/O Cost**  **(accumulated)** |
| B1 | Select rows from Item with Item\_Name = ‘Puma grey men’s’ shoes’; | 1,00,000 | 1,00,000 | 700 | 1,00,000 |
| B2 | Cartesian product (B1, Sales) | 700 + 3,000,000 | 3,000,700 | 2.1\*(10)9 | 3100700 |
| B3 | Select rows in B2 with matching Item\_ID | 2.1\*(10)9 | 2.1\*(10)9 | 700 | 2103100700 |

4.) Which plan is better to use among the above two plans in question-3? Why?

A. Plan B is better to use because you can clearly see that the “Total I/O Cost” is less when compared to Plan A. We consider the plan which I/O cost is less making it optimize.