Lab 06: Implement `Data Cubes` in Spark

[IT494, Big Data Processing, Autumn'23]

Instructor: PM Jat (pm_jat@daiict.ac.in)

In this lab, let us implement `Data Cubes` in Spark.

For dataset, use data files from Jennifer Widom discussed in lectures and are made available at https://drive.google.com/drive/folders/1GxFW26u4PF2Om0_jmQhUNcggn7WeNks9

Perform the following tasks in this lab.

1. Compute full `data cube lattice` as discussed in lectures. https://moodle.daiict.ac.in/mod/resource/view.php?id=5037 and materialize it in `parquet` data format. **For efficient querying, store every lattice node in a separate data file**.
2. Performing the following operations from the materialized lattice.
   a. Show `Item` axis
   b. Produce roll-up of `(Item, Store, Customer)`
   c. Show store-wise sales summary of `blue` `Tshirt`
   d. List all `'Tshirts'` (price <= 20) sold in `'California'` to young people (age < 25).

Submission Required:

A document in pdf format (all in one pdf file) that contains the following:

1. All source code is pasted here. Try having the code that is formatted and colored.
2. Link to “colab notebook” that contains source code and outputs of programs for all problems.

Using of markdown or latex for creating PDF would be better.