# Big Data

Tutorial #3

Isabelle Kuhlmann
2020-05-15

WeST
People and Knowledge Networks

# Outline

1) Lecture Recap
2) Exercise Discussion
3) OLAP in Spark
4) Homework Exercise

What previously happened...

# Lecture Recap

# Transactional Systems vs. Analytic Systems

| Transactional | Analytic |
|---|---|
| • Many different actions by many actors in parallel<br>• Small-sized actions spanning small fraction of data<br>• Application-oriented<br>• Current data<br>• Primary data<br>• Fast response time required<br>• Frequent changes | • Few different actions by few actors in parallel<br>• Large-sized actions spanning a lot of data<br>• Subject-oriented<br>• Historic data<br>• Aggregated data<br>• Response time: seconds, or even minutes, might be okay<br>• Supports strategic decision making |

# Data Warehousing

- Collection of data that is:
  - subject-oriented
  - integrated
  - time-variant
  - nonvolatile
- Today, we deal with more modular/agile approaches

# Metadata

- Semantic metadata
- Administrative metadata
- Schematic metadata

# Datacubes

- Multi-dimensional data representation
  - Hypercube
- Can be sparse (containing null values in facts) or dense (no null values in facts)
- Operations:
  - Slice
  - Dice
  - Roll-up
  - Drill-down
  - Pivot (rotate)

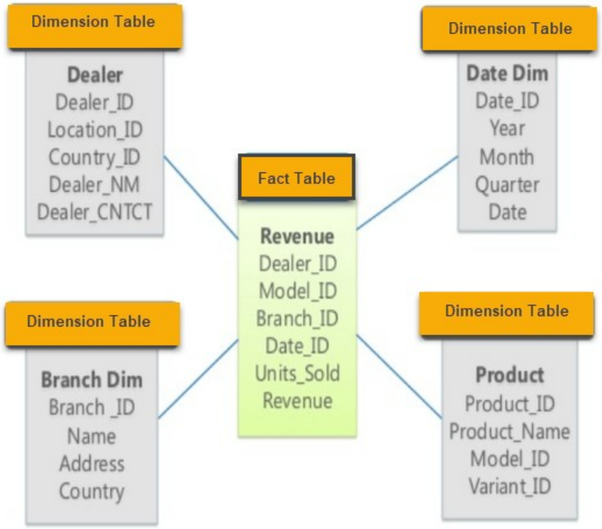
# Relational Representation

- Star schema
- Snowflake schema
- Galaxy schema

# Star Schema vs. Snowflake Schema

- Normalization of a star schema yields a snowflake schema.

**Star schema:**



**Snowflake schema:**

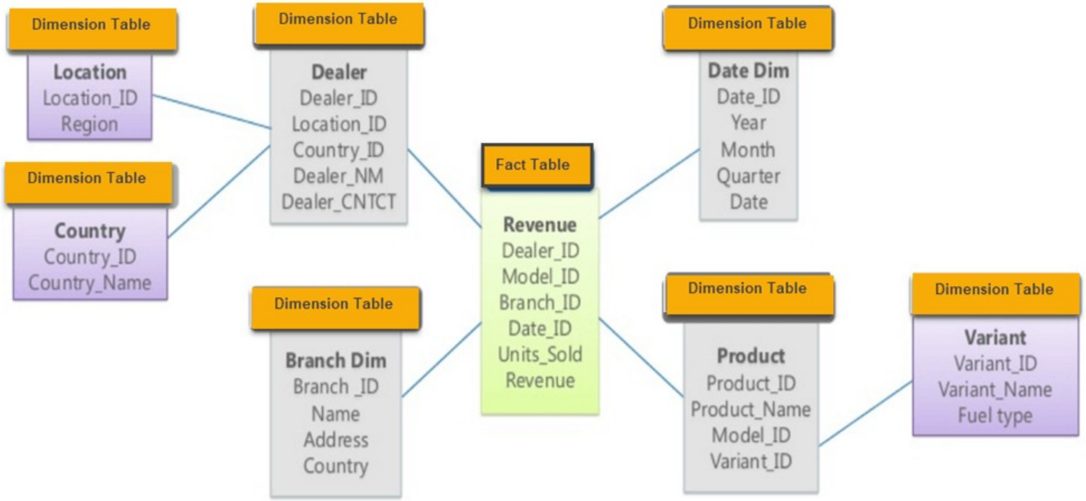


Image sources: https://www.guru99.com/images/1/022218_0758_StarandSnow1.png, https://www.guru99.com/images/1/022218_0758_StarandSnow2.png

Retail data

# Exercise Discussion

# Homework Assignment – Discussion

- Load the **Retail** dataset, it can be found under `data/retail-data/all/online-retail-dataset.csv` (Link). Answer the following questions/complete the following tasks using Spark:
  - Which item was bought most (total)? Which one was bought most in the USA?
  - Which was the lowest invoice (>0), which one the highest?
  - Add a column which displays whether an item was purchased in Germany.
  - Add a column which shows the total amount of the corresponding invoice.
  - How many German customers spent more than $10?
  - Sort the German customers with respect to their total invoice in descending order.
- Demo

*Datacubes:* Slice, Dice, Rollup, and more

# OLAP in Spark

# Slice

- Select a slice of the datacube, i.e., create a datacube with one less dimension
- Example: Demo

# Dice

- Select a subcube of your overall datacube
- Example: Demo

# Pivot

- Rotation of an axis
- Essentially yields a change in perspective
- Example: Demo

# Roll-up

- Aggregate or generalize one dimension
- In Spark, the roll-up operator basically follows a path in the lattice by always leaving out some columns
- Example: Demo

# Drill-down

- Include more specific information
  - Inverse of roll-up
- Drill-down and roll-up are effectively often a simple switch of columns (when regarding a star schema)

Now it’s your turn!

# Homework Exercise

# Exercise

- Consider the retail dataset again. Answer the following questions in two different ways: using SQL and using Spark code (DataFrame API).
  - How many orders did customers perform at which hour?
  - How frequently was each product bought in the different countries?

# Thank you for your Attention!

WeST

People and Knowledge Networks