

# Data warehouse design

Jouni Huotari & Ari Hovi

Some slides from <http://www.slideshare.net/idnats/data-warehousing-and-data-mining-presentation-725476>

13.3.2019

**jamk.fi**

# For discussion

- What is a data warehouse?
- How the term Business Intelligence relates to data warehousing?
- How data mart differs from data warehouse?
- Why OLAP (online analytical processing) is important?
- What is the basic idea behind ETL?

# What is Data Warehouse?

- Defined in many different ways, but not rigorously:
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process.” —W. H. Inmon
- Data warehousing: The process of constructing and using data warehouses

## Data Warehousing > Data Warehouse Definition

Different people have different definitions for a data warehouse. The most popular definition came from Bill Inmon, who provided the following:

**A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.**

**Subject-Oriented:** A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

**Integrated:** A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

**Time-Variant:** Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

**Non-volatile:** Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

Ralph Kimball provided a more concise definition of a data warehouse:

**A data warehouse is a copy of transaction data specifically structured for query and analysis.**

This is a functional view of a data warehouse. Kimball did not address how the data warehouse is built like Inmon did; rather he focused on the functionality of a data warehouse.

# Basic design principles for data warehouses

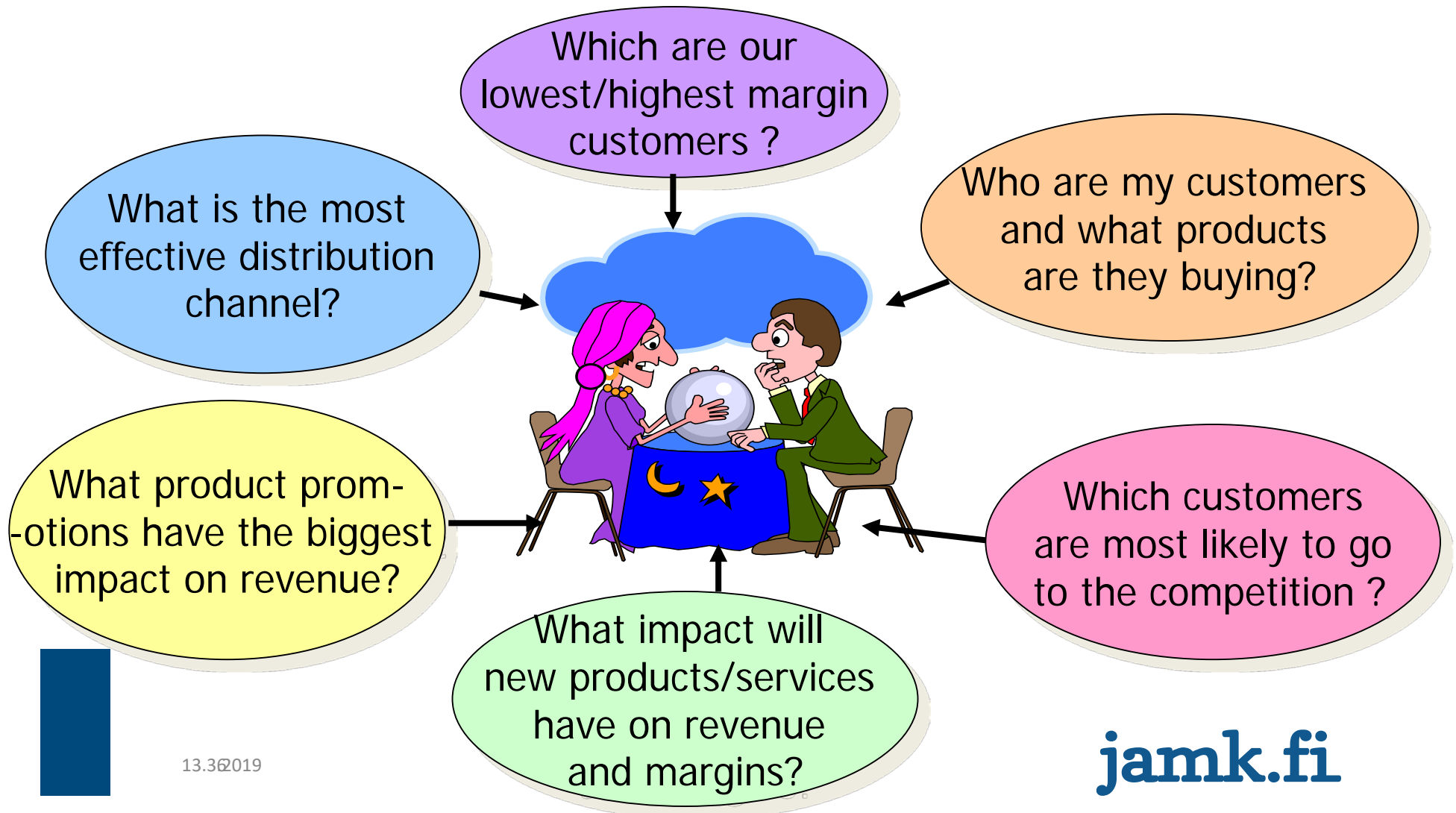
To support fast summary queries, analysis, and reporting

- This is often difficult in operative databases, but some solutions exist
- *can you mention any solutions?*

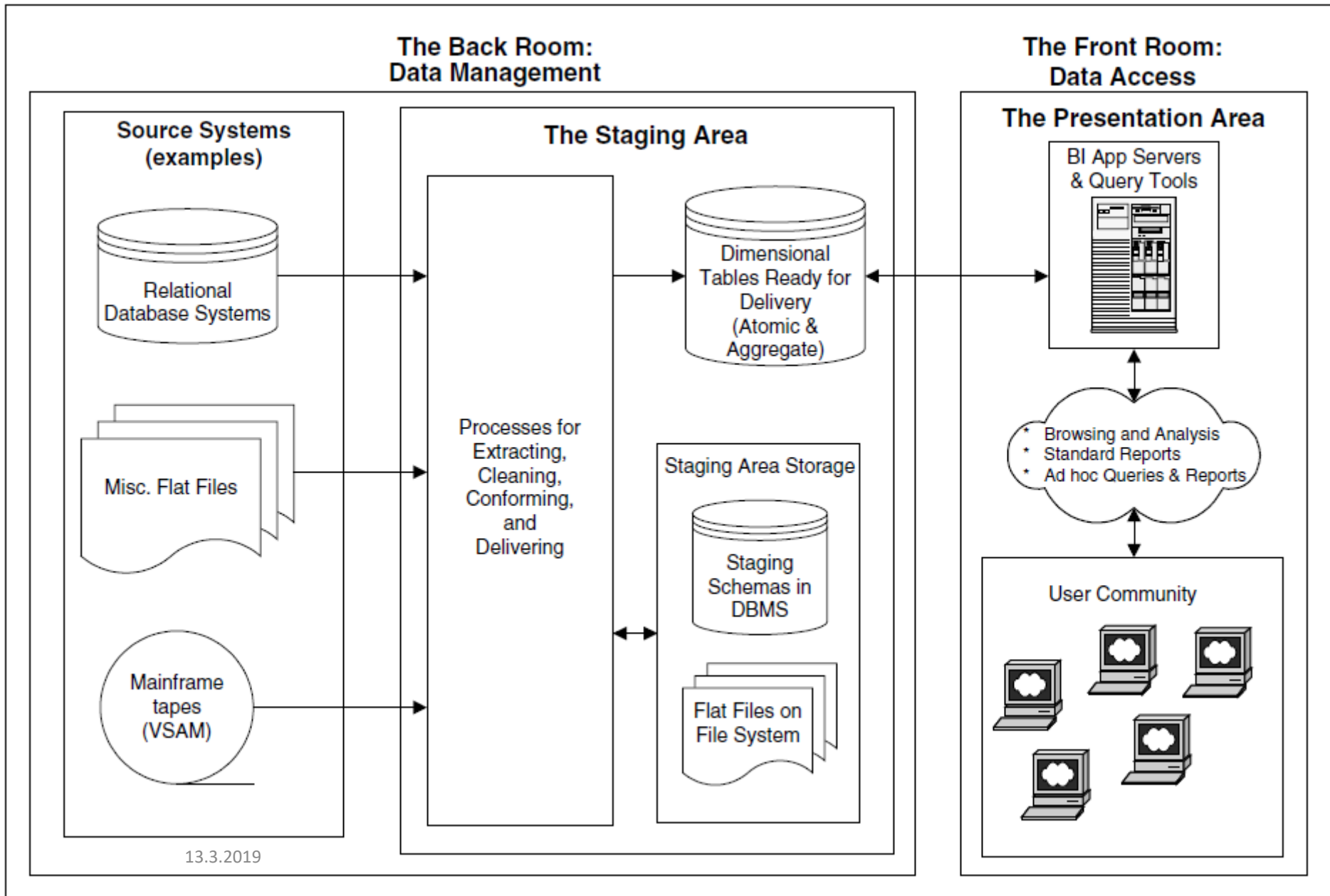
It is important to

- Maintain history for seeing trends etc.
- Clean up and make data consistent from different data sources
- Make the structure clear and understandable

# Example: a producer wants to know...



# back room and front room of a data warehouse



# Data Warehouse vs. Operational DBMS

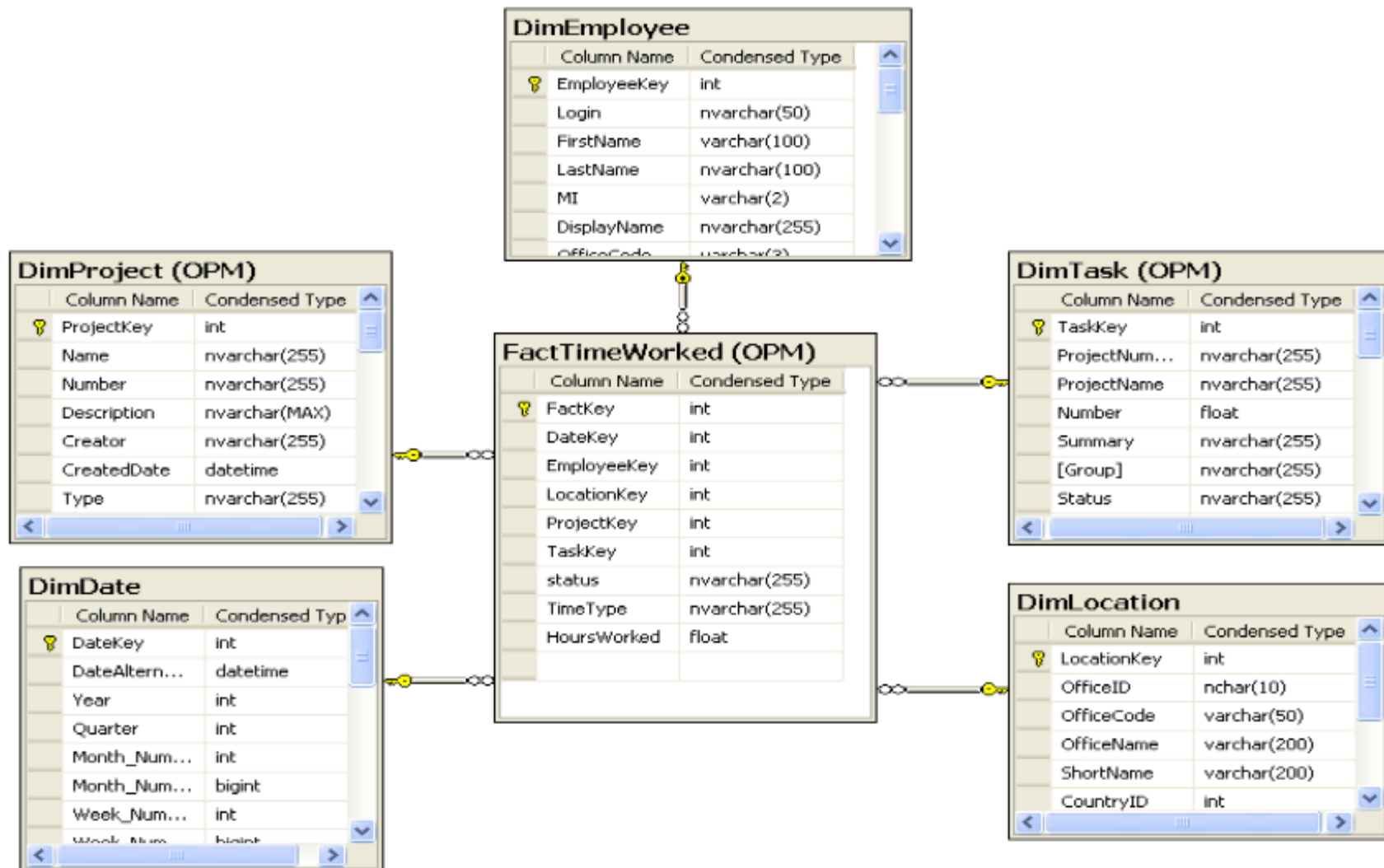
- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries



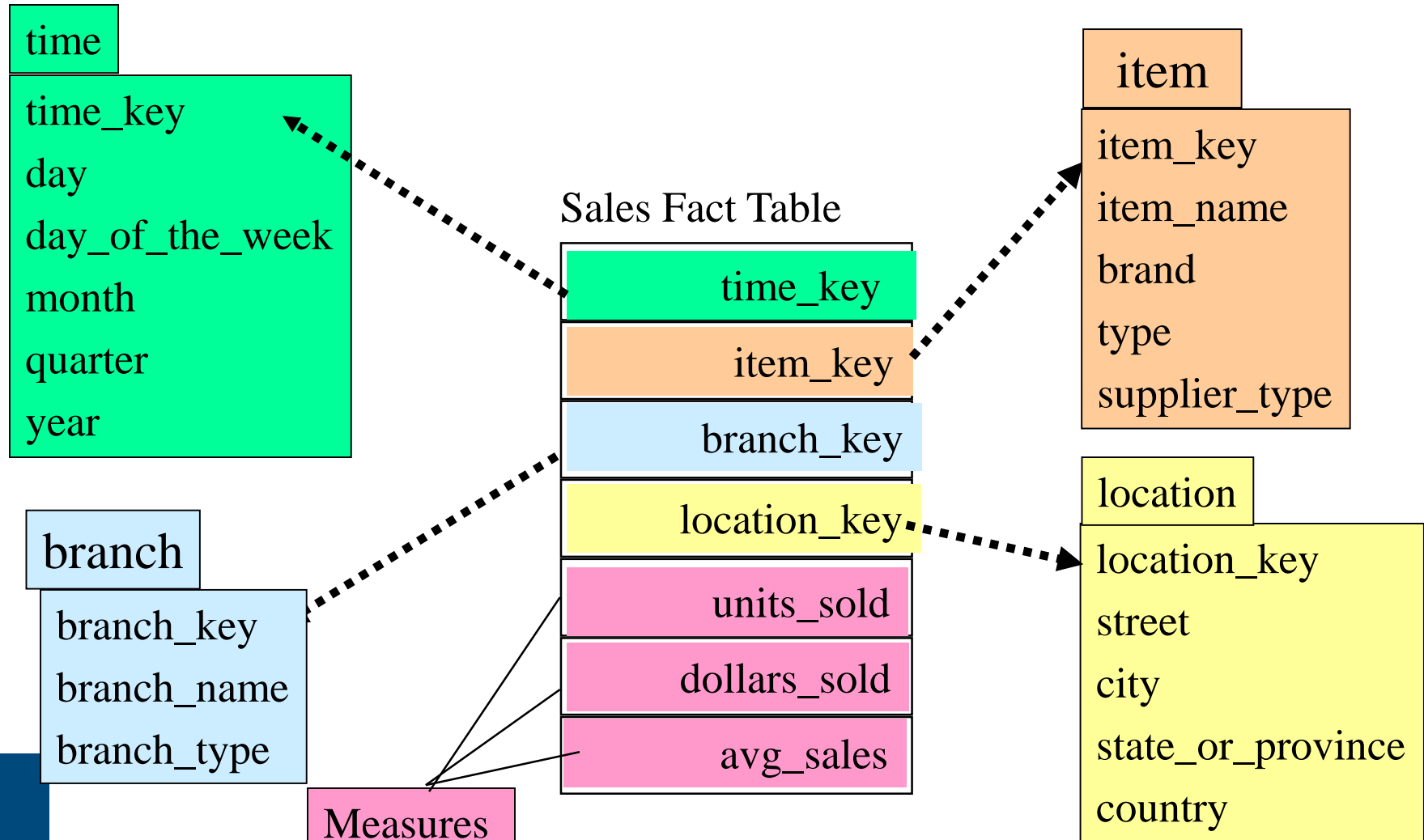
# Conceptual Modeling of Data Warehouses

- **Star schema:** A fact table in the middle connected to a set of dimension tables
- **Snowflake schema:** A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
- **Fact constellations:** Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation
- **Dimensions** describe who, what, when, where and why for the facts.

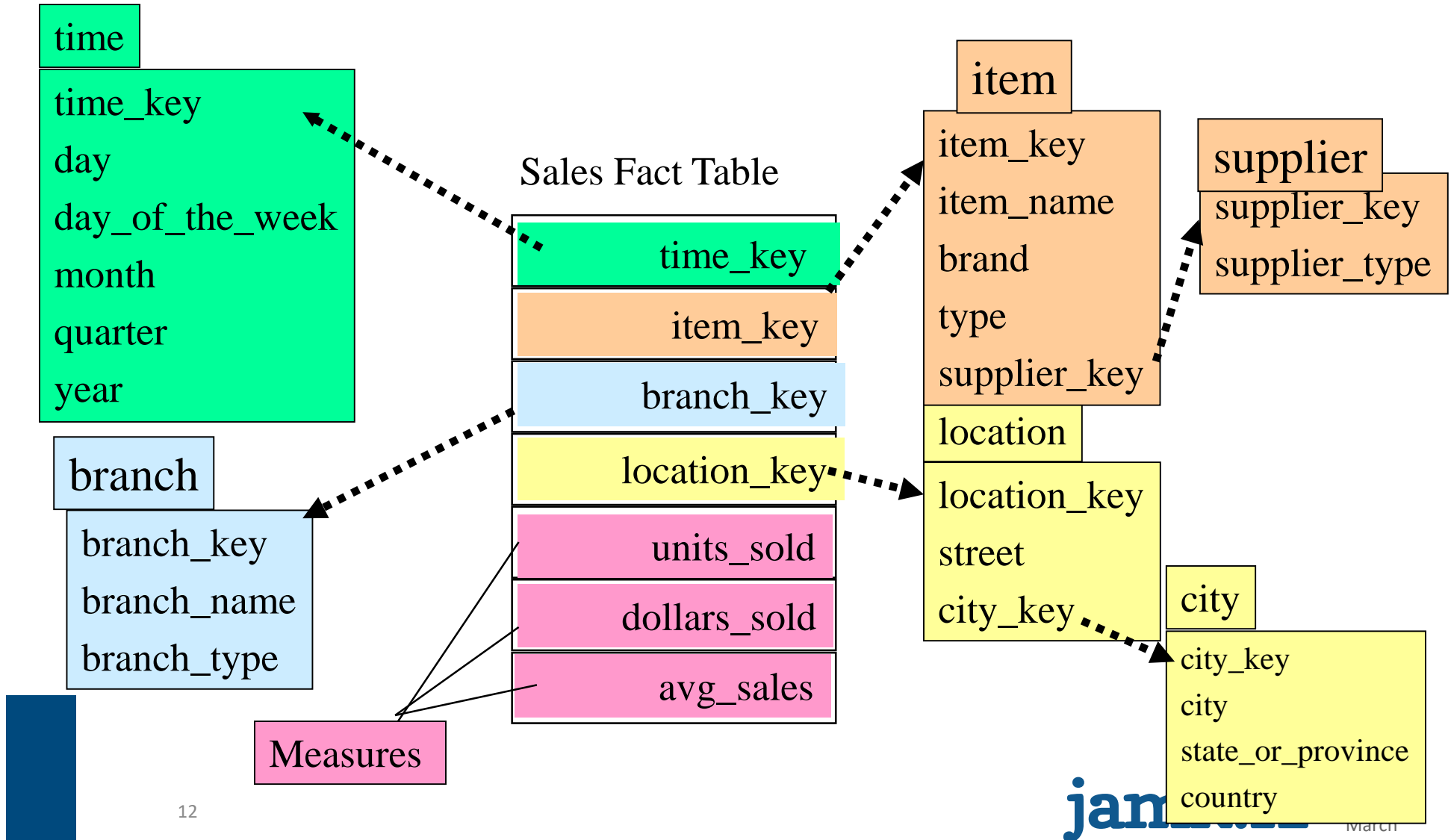
# Example of Star Schema



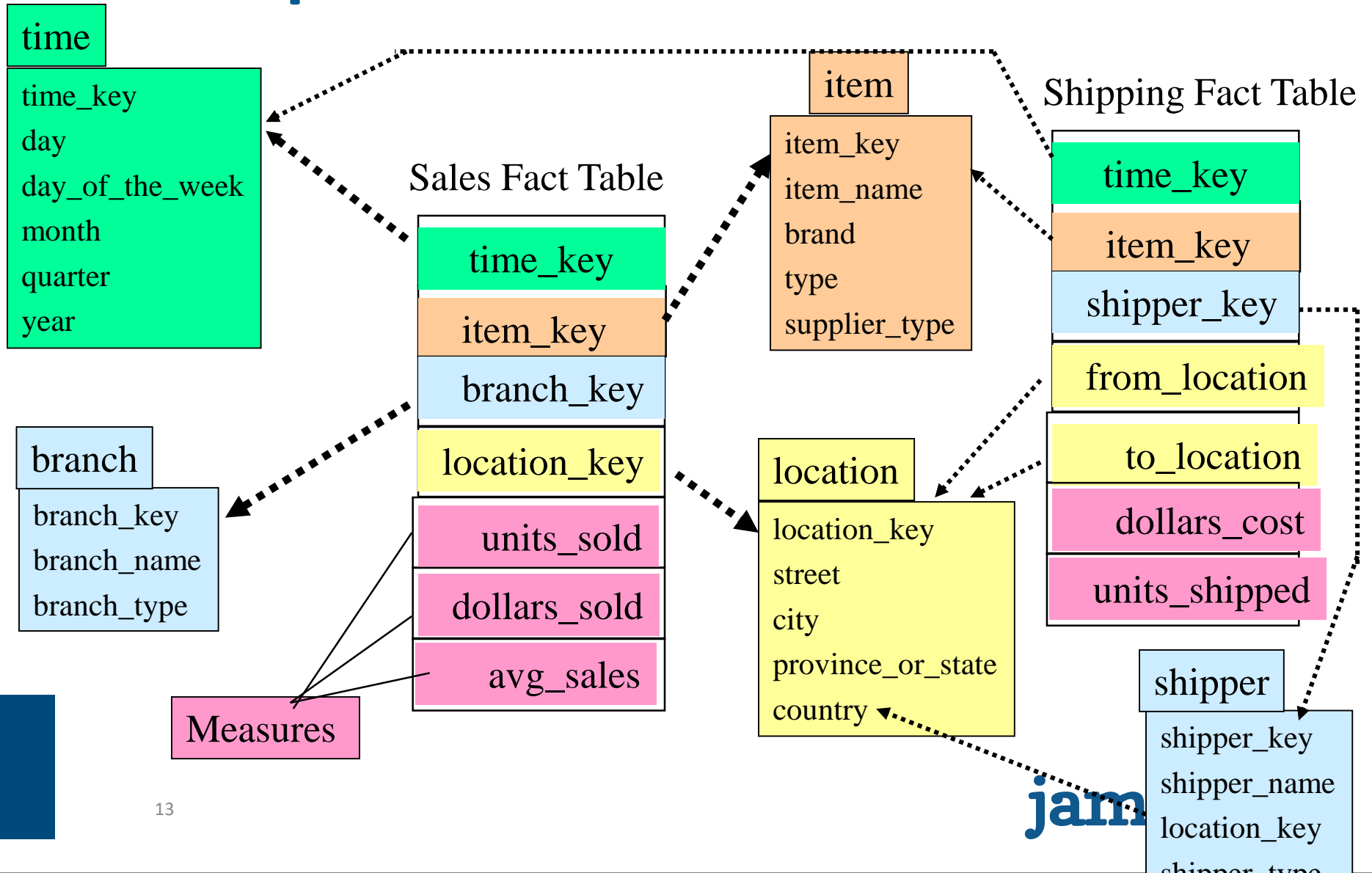
# Another example of Star Schema



# Example of Snowflake Schema

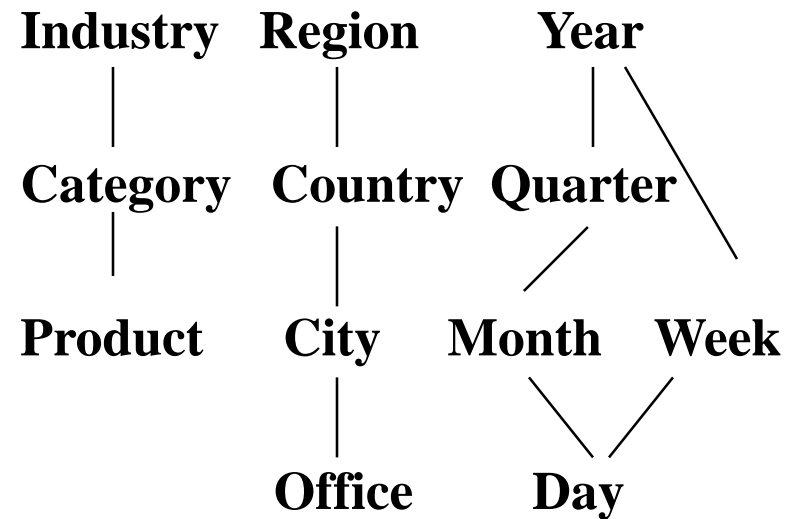
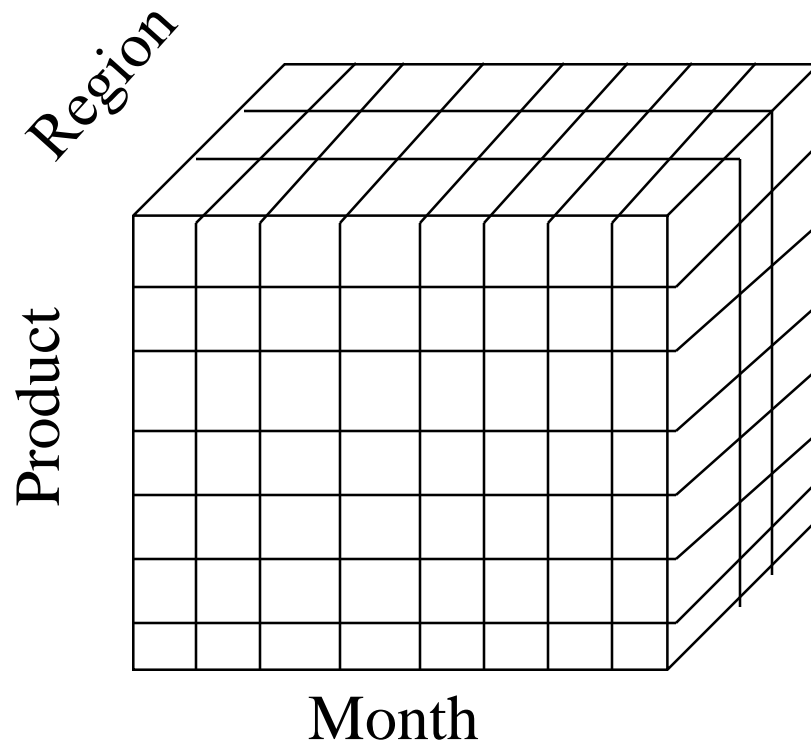


# Example of Fact Constellation

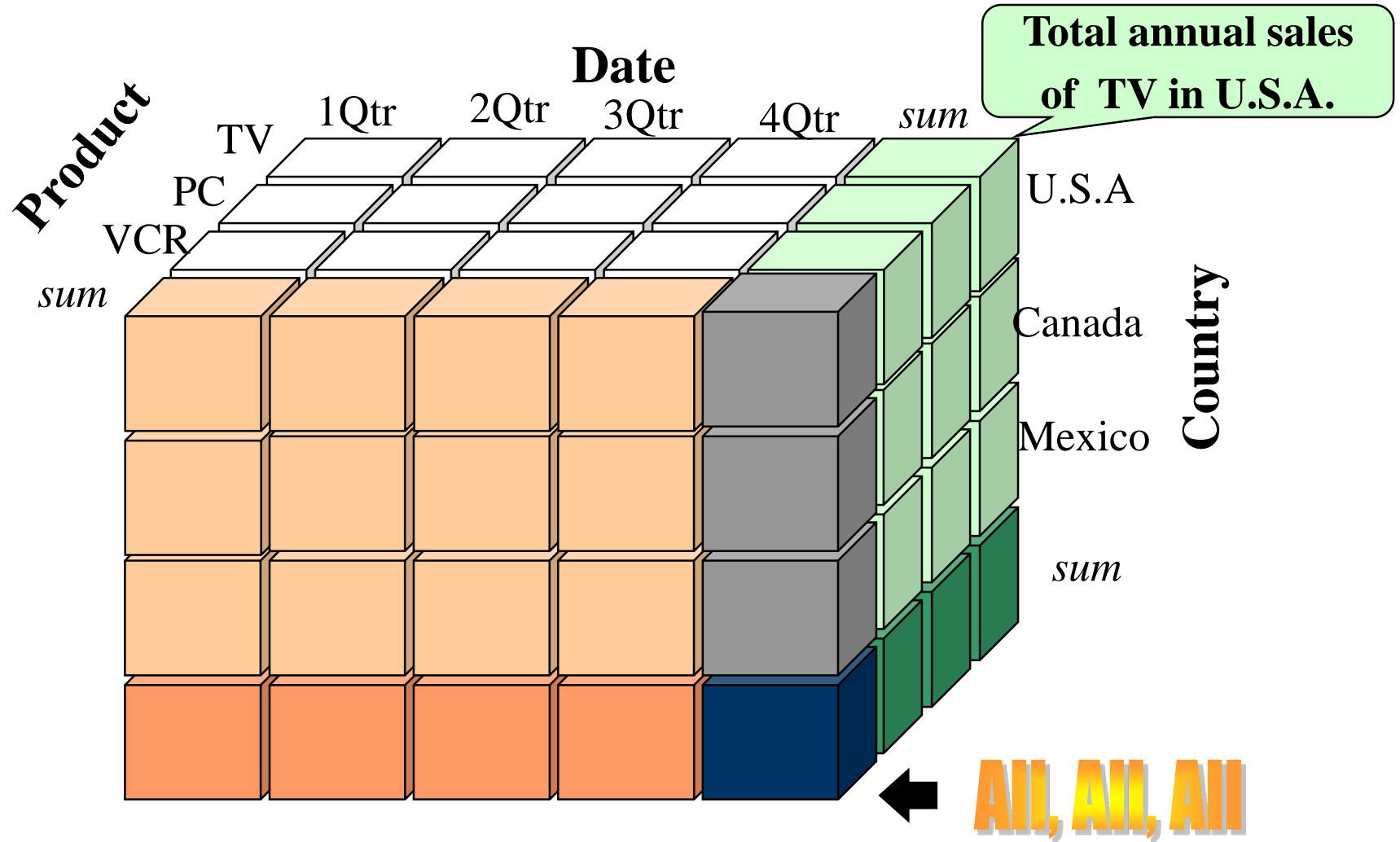


# Multidimensional Data

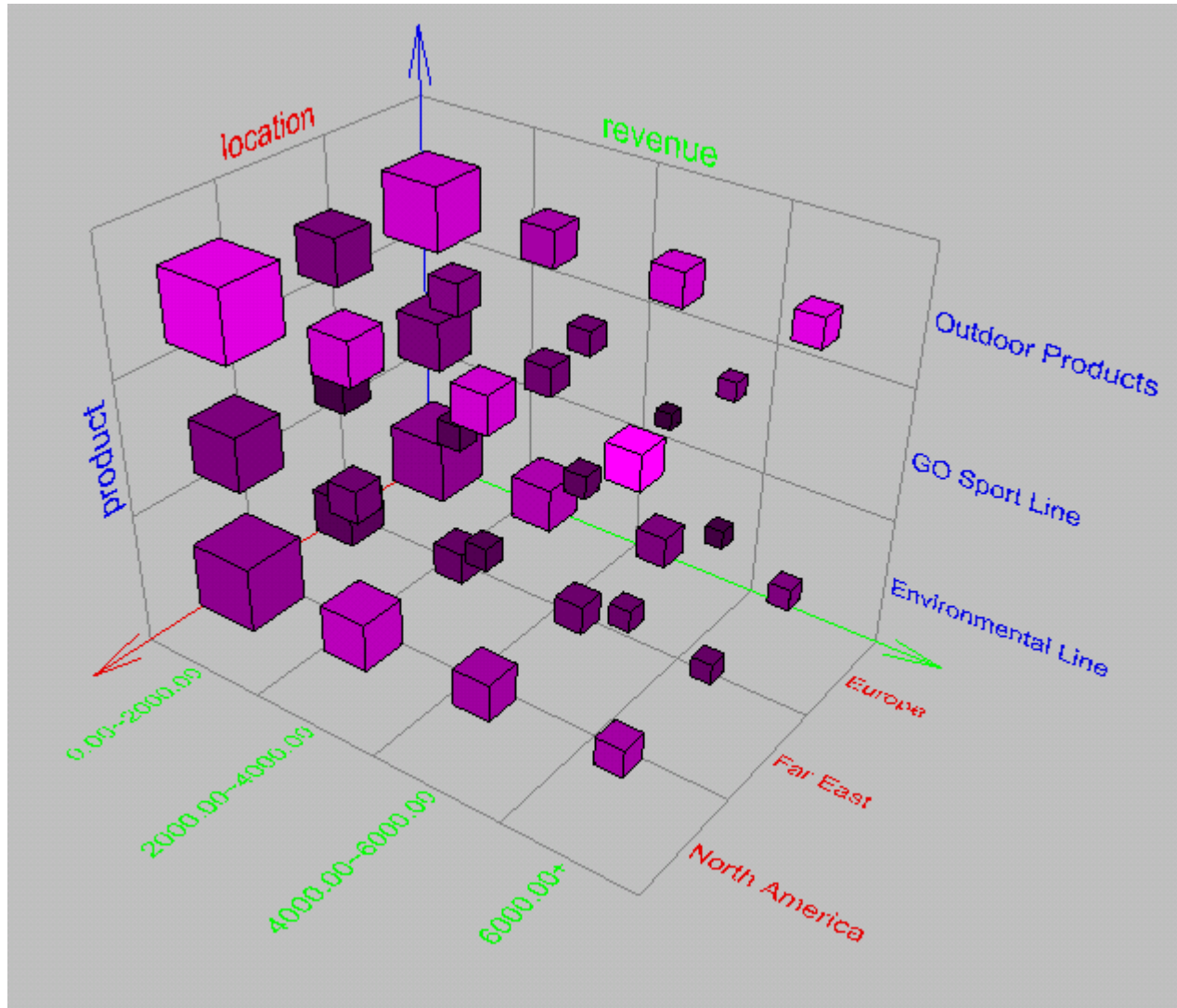
- Sales volume as a function of product, month, and region
- **Dimensions: Product, Location, Time**
- **Hierarchical summarization paths**



# A Sample Data Cube



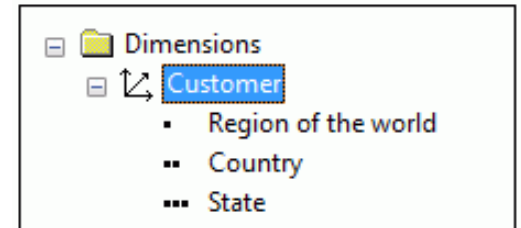
# Browsing a Data Cube



- Visualization
- OLAP capabilities
- Interactive manipulation



# Exercises



- Install <http://www.olapcube.com/> to your virtual machine and play with the dimensions, create a cube and examine the result (dashboard)
- <http://www.olapcube.com/help/writer/panorama/>
- <http://www.pentaho.com/testdrive>
- [http://www.databaseanswers.org/downloads/Data Warehousing by Example.pdf](http://www.databaseanswers.org/downloads/Data_Warehousing_by_Example.pdf)
- Example (real) data: <http://www.gapminder.org/data/>

# What is Business Intelligence (BI)?

- BI refers to skills, technologies, applications and practices used to help a business acquire a better understanding of its commercial context (Wikipedia)
- The goal is to gain insight into the business by bringing together data, formatting it in a way that enables better analysis, and then providing tools that give users power—not just to examine and explore the data, but to quickly understand it. (Business Intelligence with Microsoft Office PerformancePoint Server)

## More information from wikipedia:

- [http://en.wikipedia.org/wiki/Data\\_warehouse](http://en.wikipedia.org/wiki/Data_warehouse)
- [http://en.wikipedia.org/wiki/Data\\_mart](http://en.wikipedia.org/wiki/Data_mart)
- [http://en.wikipedia.org/wiki/Star\\_schema](http://en.wikipedia.org/wiki/Star_schema)
- [http://en.wikipedia.org/wiki/Snowflake\\_schema](http://en.wikipedia.org/wiki/Snowflake_schema)
- <http://en.wikipedia.org/wiki/OLAP>

# Example

Panoply

<https://panoply.io/>



- Home
- Data Sources
- Tables
- Analyze
- Query Log
- Jobs
- Alerts
- Teams
- Connect

Choose source type Search

Most Popular (10) >

- Sample Source (1)
- Files & Services (9)
- APIs (67)
- Databases (12)

Sample Data Use this data for a test run

- File Upload
- Amazon S3
- Amazon SQS
- Facebook P
- Google Analytics
- Mongo DB
- MySQL
- Panoply SD
- Postgres

Something's missing? Request

Database Created Collect Data Query It Connect

Home

Data Sources

Tables

Analyze

Query Log

Jobs

Alerts

Teams

Connect

## 1 Data Sources

🔒 Source credentials and parameters are encrypted

[Add D](#)

### Sample Data

**File name** pokemon.csv

Supported formats: CSV & TSV, JSON & JSON-lines, XLS, Web-distribution logs, query-string logs, gzip, zip, tar [and more...](#)

**Destination**

Name of the target table where you wish to save data

**Advanced**[Show](#)[Collect](#)[✔ Database Created](#)[Collect Data](#)[Query It](#)[Connect Any B](#)

# 1 Data Sources

 Source credentials and parameters are encrypted

Add Data Source



Sample Data

running...



File name

Choose File

pokemon.csv

Supported formats: CSV & TSV, JSON & JSON-lines, XLS, Web-distribution logs, query-string logs, gzip, zip, tar and more...

Destination

pokemon

Name of the target table where you wish to save the data

In the meantime - create another data source?



Database Created



Collect Data

Query It

Connect Any BI Tool

6	060ad92489947d410d8974740...	2019-03-12T15:43:29.205Z	450	Field
---	------------------------------	--------------------------	-----	-------

## Metadata

[Show Hidden \(3\)](#)

## Connection Details

Text	type_1	×
Text	id	
X <sub>2</sub> Number	total	×
Text	egg_group_1	×
Text	islegendary	×
Text	color	×
X <sub>2</sub> Number	attack	×

**HOST** db.panoply.io

**DATABASE** ttow0110

**PORT** 5439

**USER** huojo.jamk@gmail.com

**PASSWORD**(same as your Pa

**DRIVER** Amazon Redshift o

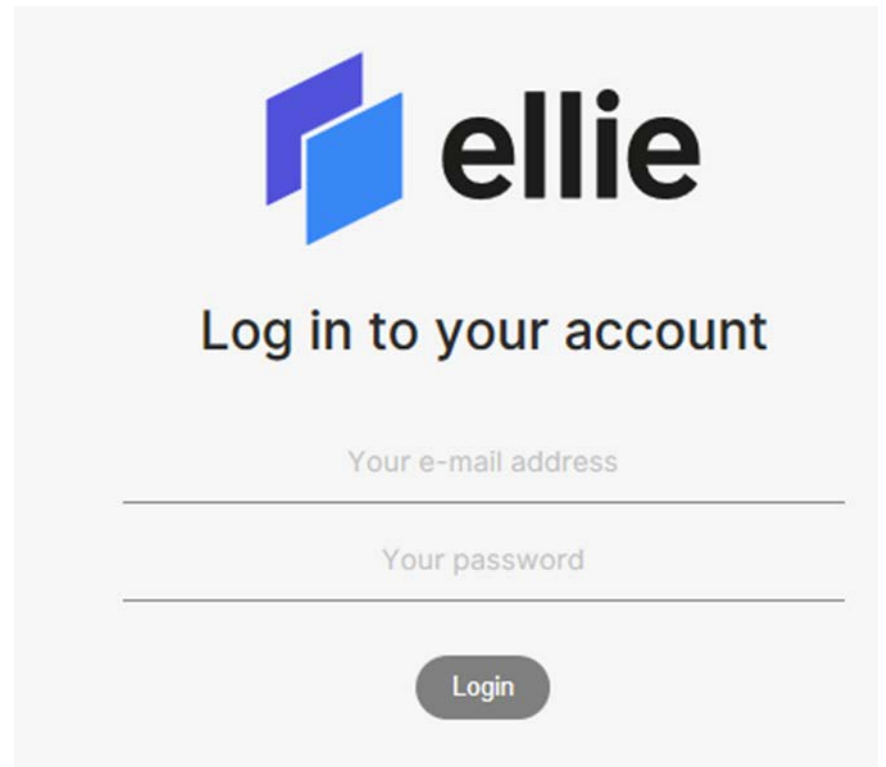


```
SELECT * FROM "public"."pokemon" LIMIT 10
```

id	__updateTime	total	en
006f52e9102a8d3be2fe5614f42...	2019-03-12T15:43:29.205Z	390	Bug
01386bd6d8e091c2ab4c7c7de6...	2019-03-12T15:43:29.205Z	495	Water_3
02522a2b2726fb0a03bb19f2d8...	2019-03-12T15:43:29.205Z	525	Field
0353ab4cbcd5beae847a7ff6e2...	2019-03-12T15:43:29.205Z	510	Field
045117b0e0a11a242b9765e79...	2019-03-12T15:43:29.205Z	365	Monster
0584ce565c824b7b7f50282d9a...	2019-03-12T15:43:29.205Z	680	Undiscovere
06409663226af2f3114485aa4e...	2019-03-12T15:43:29.205Z	314	Monster
072b030ba126b2f4b2374f342b...	2019-03-12T15:43:29.205Z	300	Water_1
077e29b11be80ab57e1a2ecab...	2019-03-12T15:43:29.205Z	680	Undiscovere
07e1cd7dca89a1678042477183...	2019-03-12T15:43:29.205Z	450	Water_2

Database Created Collect Data Query It

## Another example: Ellie



The image shows a login form for 'ellie'. At the top left is the logo, which consists of two overlapping squares, one purple and one blue. To the right of the logo is the word 'ellie' in a bold, lowercase, sans-serif font. Below the logo and name is the text 'Log in to your account'. Underneath this are two input fields: the first is labeled 'Your e-mail address' and the second is labeled 'Your password'. Both fields are represented by horizontal lines. At the bottom of the form is a rounded rectangular button with the text 'Login' inside.

<https://jamk.ellie.fi>



**jamk.fi**