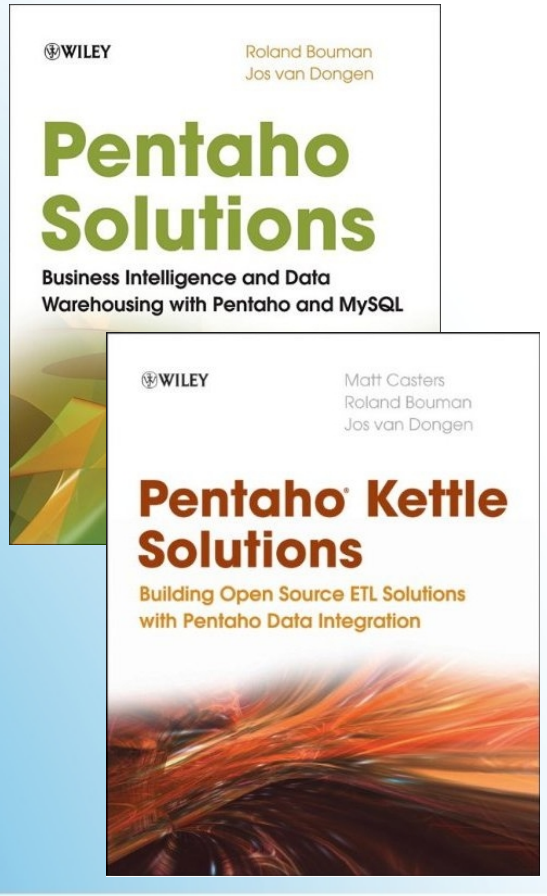
A Data Warehousing and Business Intelligence Tutorial

Starring Sakila

Welcome!

Matt Casters
Chief Data Integration, Pentaho
http://www.ibridge.be/
@mattcasters

Roland Bouman
Software Engineer, Pentaho
http://rpbouman.blogspot.com/
@rolandbouman



Starring Sakila

Pentaho

- Commercial Open Source Business Intelligence
  - Full BI suite since 2005
- Projects: Kettle (DI & ETL), Jfree (Reporting), Mondrian (OLAP), Weka (Data Mining)
- Community: CDF (Dashboarding), Saiku (OLAP)
- Recent: Focus on “Big Data”, esp. Hadoop
- http://www.pentaho.com
- http://sourceforge.net/projects/pentaho/

- Business Intelligence
- Data Warehousing
- Anatomy of a Data Warehouse
- Physical Implementation
- Sakila – a Star is Born
- Filling the Data Warehouse
- Presenting the Data - BI Applications

Agenda

Part I:

Business Intelligence

Starring Sakila

- Skills, technologies, applications and practices to acquire a better understanding of the commercial context of your business
- Turning data into information useful for business users
  - Management Information
  - Decision Support

Business Intelligence

months, years:

“Should we become an appliance vendor instead of delivering software solutions”

weeks, months:

“In what region should we open a new store?”

days, weeks:

“Who's available for tomorrow's shift”

Data mining

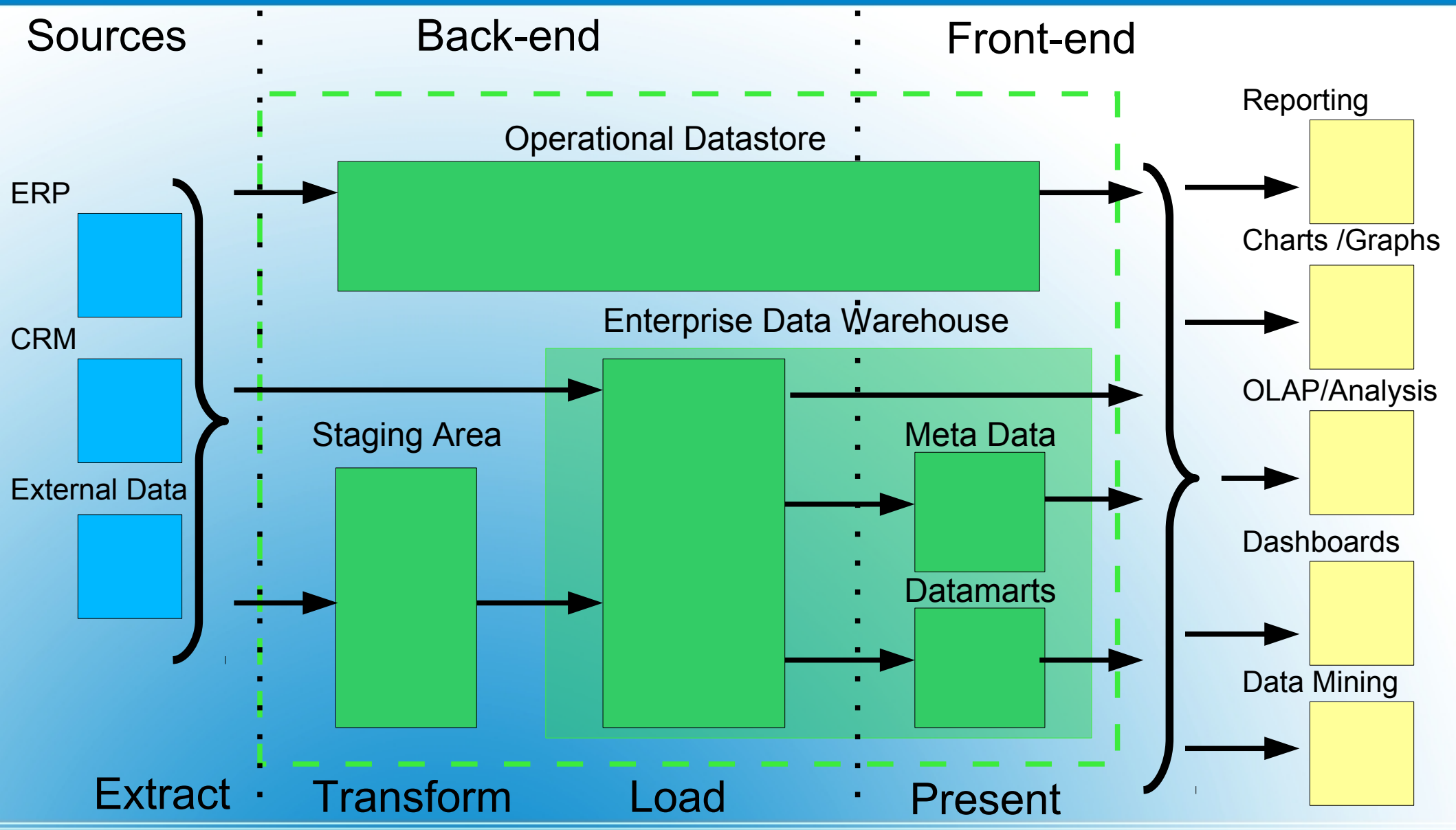
OLAP/Analysis

Reporting

Business Intelligence Scope

- Front end Applications:
  - Reports
  - Charts and Graphs
  - OLAP Pivot tables
  - Data Mining
  - Dashboards
- Back end Infrastructure
  - Data Integration
  - Data Warehouse
  - Data Mart
  - Metadata
  - (ROLAP) Cube

Functional Parts of a Business Intelligence Solution



High Level BI Architecture

Part II:

Data Warehousing

Starring Sakila

What is a Data Warehouse?

- A database designed to support Business Intelligence Applications
- Different requirements as compared to Operational Applications
- Analytic Database Systems (ADBMS)
  - MySQL: Infobright, InfiniDB
  - LucidDB, MonetDB

What is a Data Warehouse?

- Ultimately, it's just a Relational Database
  - Tables, Columns, ...
- Designed for Business Intelligence applications
  - Ease of use
  - Performance
- Data from various source systems
  - Integration, Standardization, Data cleaning
- Add and maintain history
  - Corporate memory

- A database designed to support BI applications
- BI applications (OLAP) differ from Operational applications (OLTP)
  - OLTP: Online Transaction Processing
  - OLAP: Online Analytical Processing
- Differences:
  - Applications, Data Processing, Data Model

What is a Data Warehouse?

OLTP vs OLAP: Application Characterization

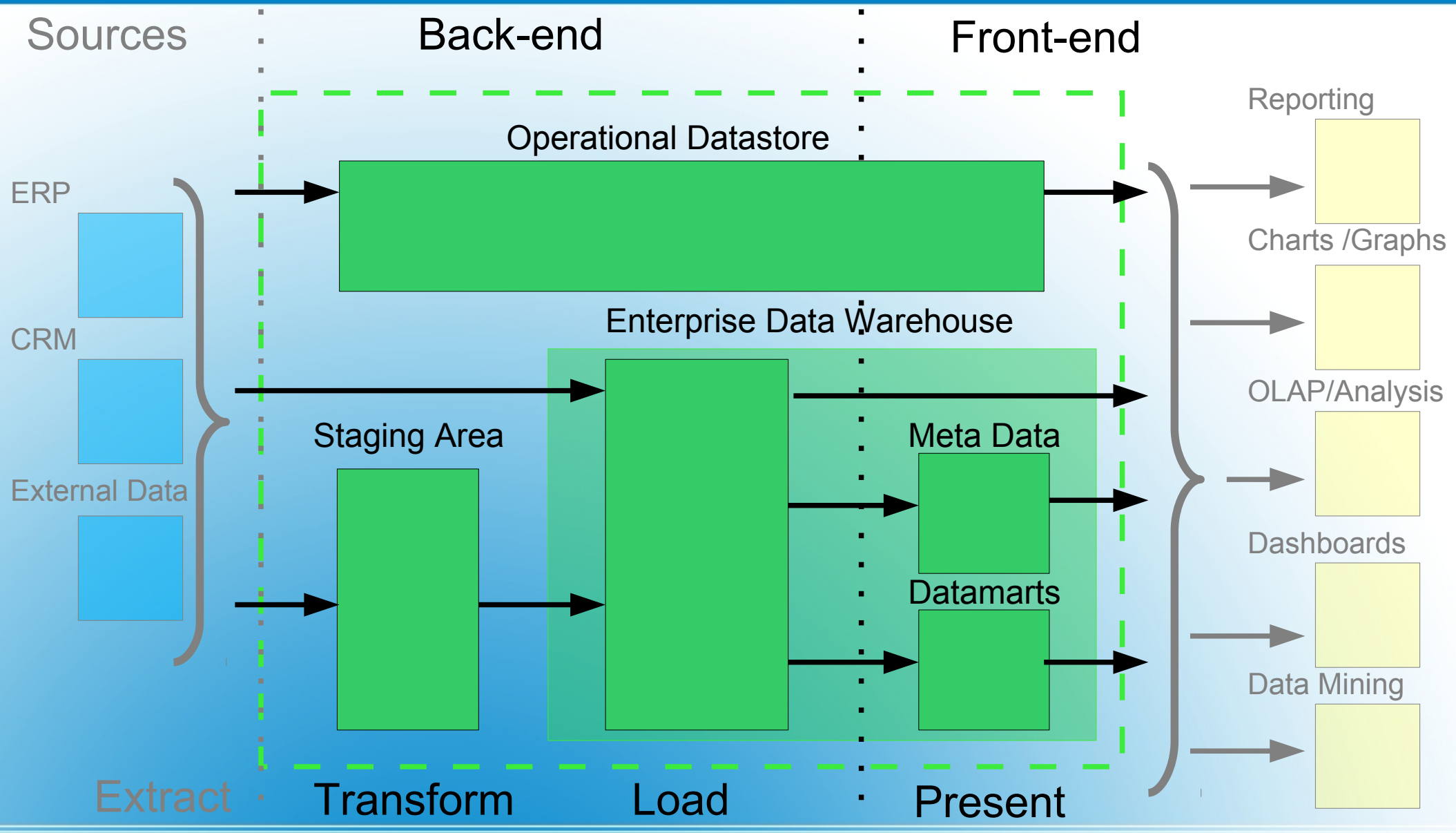
- OLTP
  - Operational
  - 'Always' on
  - All kinds of users
  - Many users
  - Directly supports business process
  - Keep a Record of Current status
- OLAP
  - Tactical, Strategic
  - Periodically Available
  - Managers, Directors
  - Few(er) users
  - Redesign Business Process
  - Decision support, long-term planning
  - Maintains a history

OLTP vs OLAP: Data Processing

- OLTP
  - Transactions
  - Subject Oriented
  - Add, Modify, Remove single rows
  - Human data entry
  - Queries for small sets of rows with all their details
  - Standard queries
- OLAP
  - Groups
  - Aspect Oriented
  - Bulk load, rarely modify, never remove
  - Automated ETL jobs
  - Scan large sets to return aggregates over arbitrary groups
  - Ad-hoc queries

OLTP vs OLAP: Data Model

- OLTP
  - Entity-Relationship model
  - Entities, Attributes, Relationships
  - Foreign key constraints
  - Indexes to increase performance
  - Normalized to 3NF or BCNF
- OLAP
  - Dimensional model
  - Facts, Dimensions, Hierarchies
  - Ref. integrity ensured in loading process
  - Scans on Fact table obliterates indexes
  - Denormalized Dimensions (<= 1NF)



High Level BI Architecture

Part III:

Dimensional Model

Starring Sakila

What is the Dimensional Model?

- An aspect-oriented logical data model optimized for querying and data presentation
- Divides data in two kinds:
  - Facts
  - Dimensions

- Facts
  - Measures/Metrics of a Business Process
  - Examples: Cost, Units Sold, Profit
- Dimensions
  - Context of Business Process
  - Who? What? Where? When? Why?
  - Navigate Facts: Selection, Rollup, Drilldown
  - Provide and maintain history

The Dimensional Model

Date Dimension →		2008 Q4			
Location Dimension ↓		All Months	October	November	December
All locations		$ 3850	$ 1000	$ 1350	$ 1500
America	All America	$ 2050	$ 500	$ 750	$ 800
	North	$ 1275	$ 300	$ 500	$ 475
	South	$ 775	$ 200	$ 250	$ 325
Europe	All Europe	$ 1800	$ 500	$ 600	$ 700
	East	$ 800	$ 250	$ 250	$ 300
	West	$ 1000	$ 250	$ 350	$ 400

Dimensional Data Presentation

The Dimensional Model: Facts

- Fact table structure:
  - Several measures
  - Keys to dimension tables
- Measures:
  - Usually numeric, Additive, Semi-additive
  - Sometimes pre-calculated
- Rapidly growing!
  - Millions, Billions of rows (Terabytes)

- Dimension table structure:
  - Surrogate key and descriptive text attributes
- Relatively few rows
  - Exception: Customer 'Monster' dimension
- Relatively static
  - Exception: Slowly changing dimensions
- Used to navigate through fact data
  - Hierarchies

The Dimensional Model: Dimensions

- Selection (Filter)
- Navigation: Attributes organized in Hierarchies
  - Date dimension examples:
    - Year, Quarter, Month, Day
    - Year, Week, Day
- Groupings for Aggregation
  - 'Roll up', 'Drill Down'
  - 'Slice and Dice'

The Dimensional Model: Navigating data with Dimensions

The Dimensional Model: Maintaining History

- Fact table usually links to a date dimension
- Dimensions maintain their own history
  - Slowly changing dimensions
- Type I Overwrite (no history)
- Type II
  - History kept in rows (versioning)
- Type III
  - History kept in columns

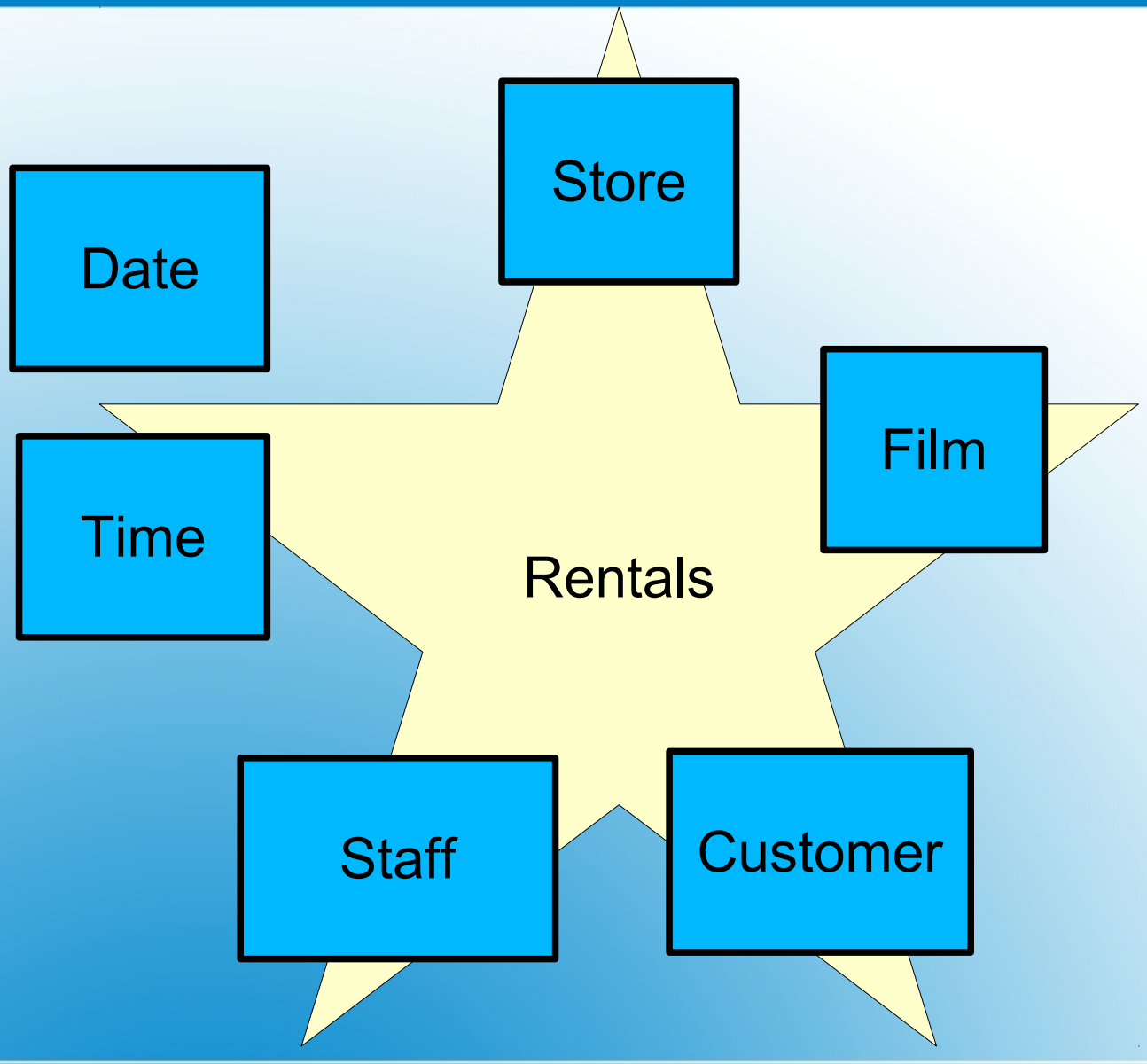
Part V:

Physical Implementation

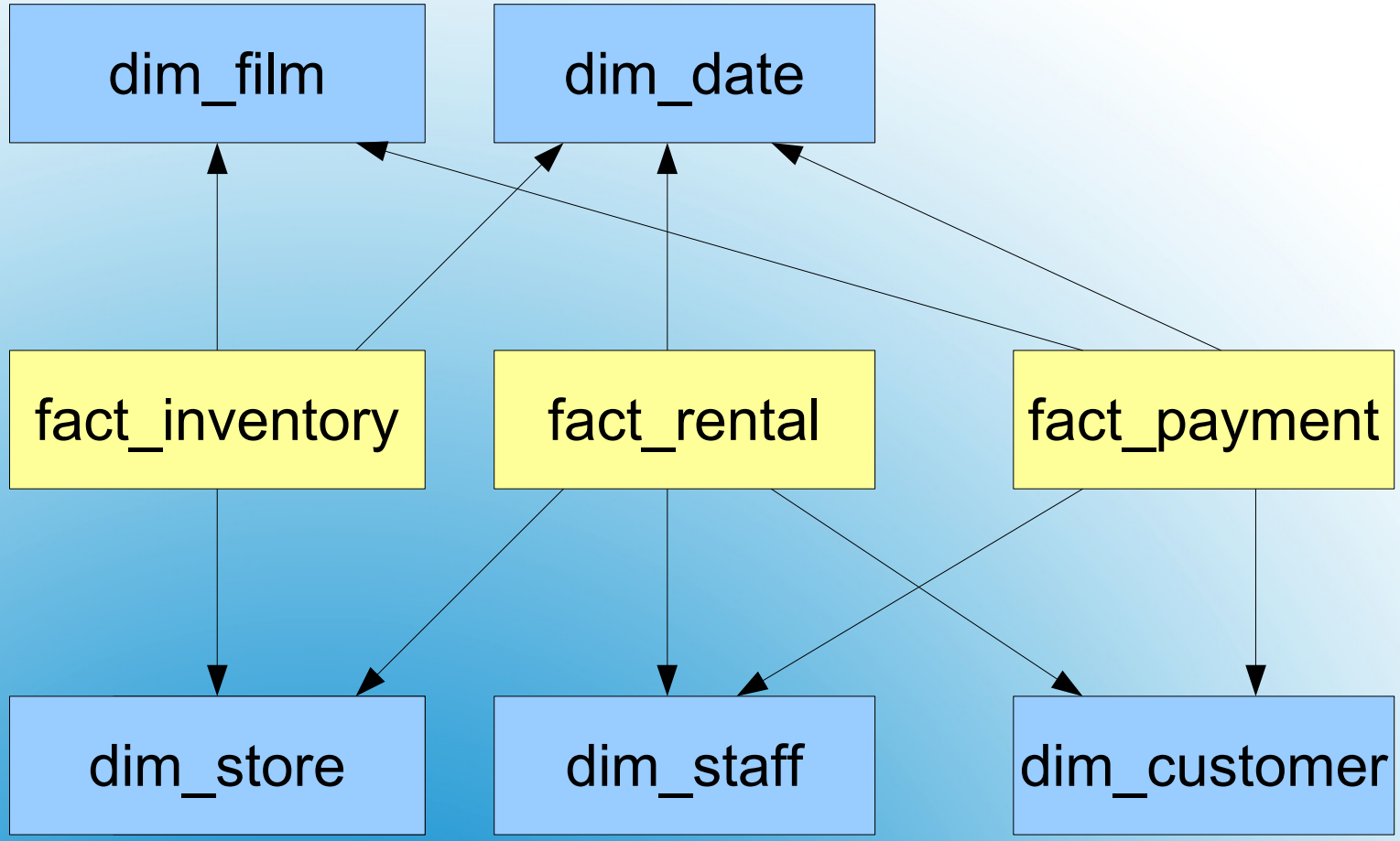
Starring Sakila

- Related metrics stored in a Fact table
- Fact table references relevant dimensions
- Each Dimension stored in a Dimension Table
- Dimension tables shared by multiple fact tables

Dimensional Model Implementation: Star Schema



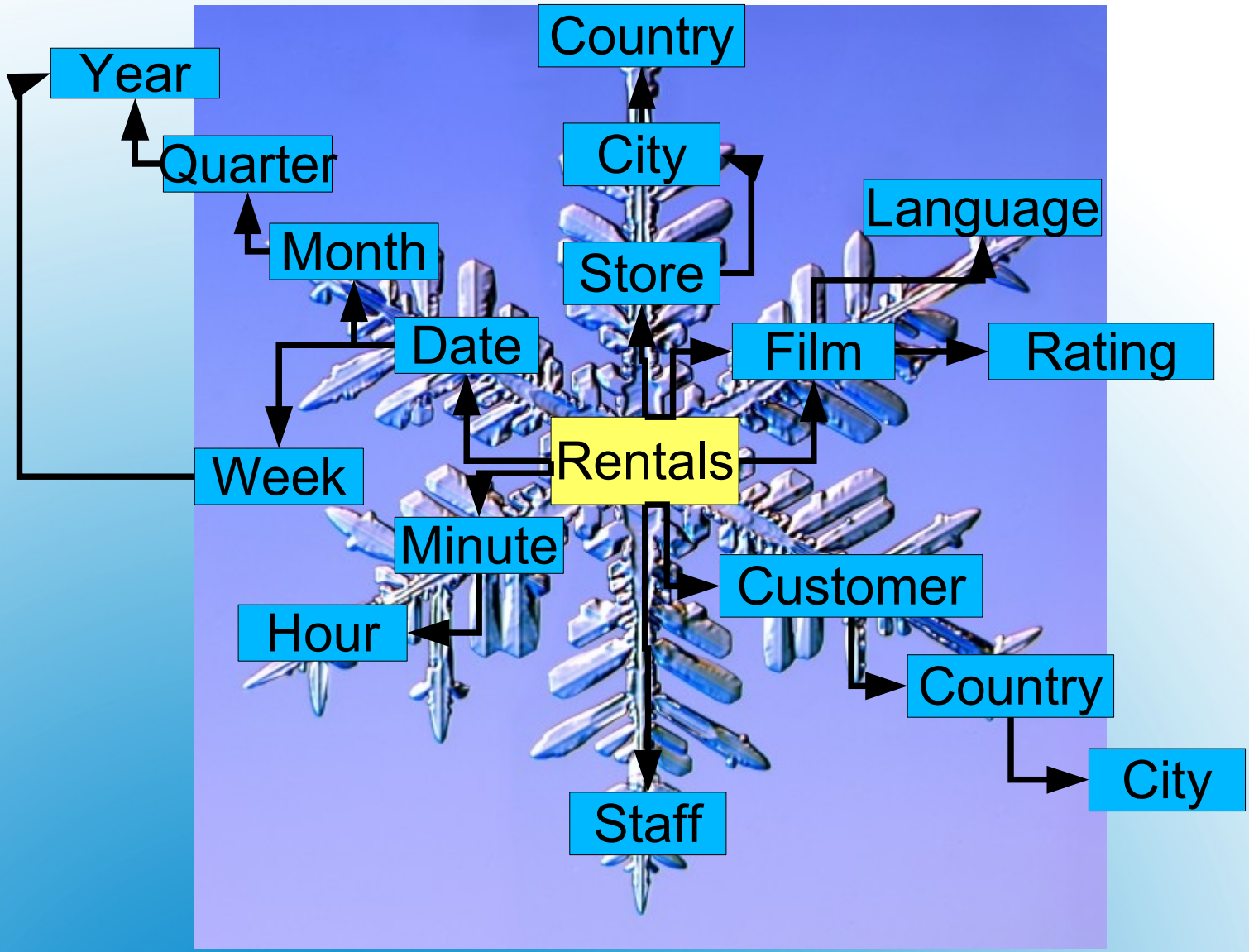
Star Schema example: Sakila Rentals



Star Schema example: Sakila Rentals

Stars versus Snowflakes

- Star schema is 'just' an implementation
  - Optimized for simplicity
  - Optimized for performance (?)
  - Heavily denormalized dimensions
- There is an alternative: Snowflake
  - Normalized dimensions



Snow Flake example: Sakila Rentals

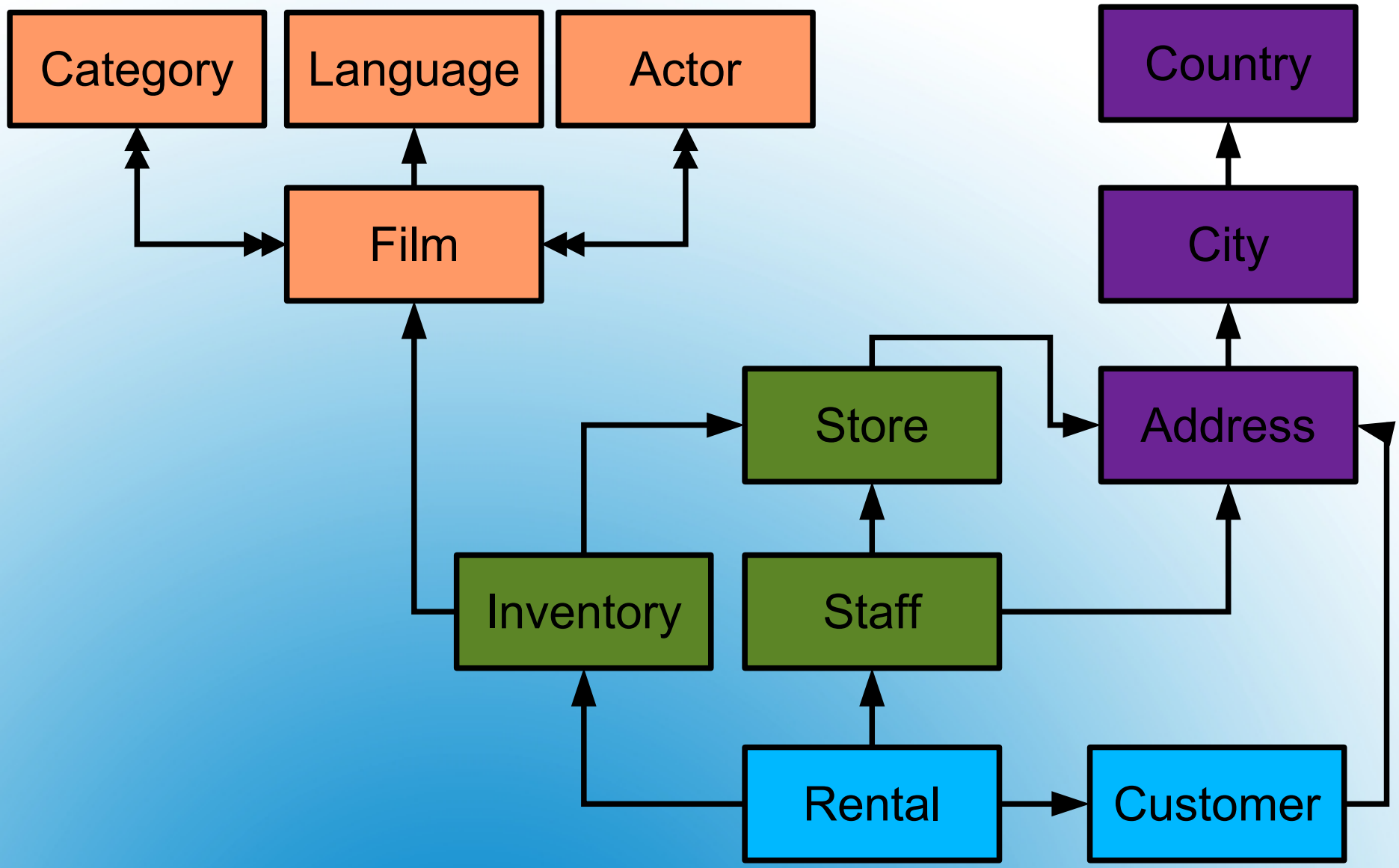
Part V:

A Star is Born

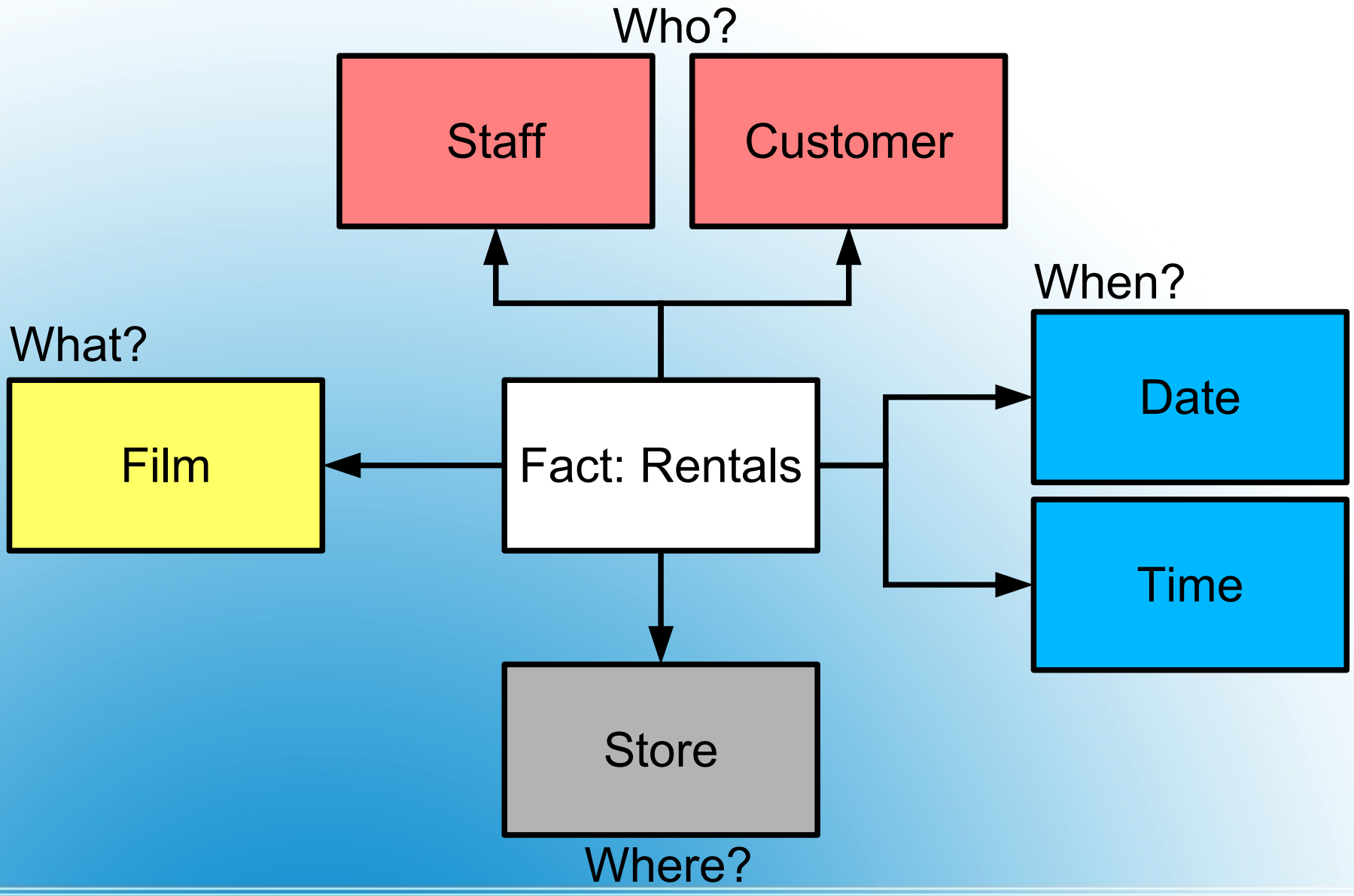
Starring Sakila

- MySQL Sample Database
  - http://dev.mysql.com/doc/sakila/en/sakila.html
- DVD rental business
  - Overly simplified database schema
- Typical OLTP database

Dimensional Model example



3NF Source schema: Sakila Rentals



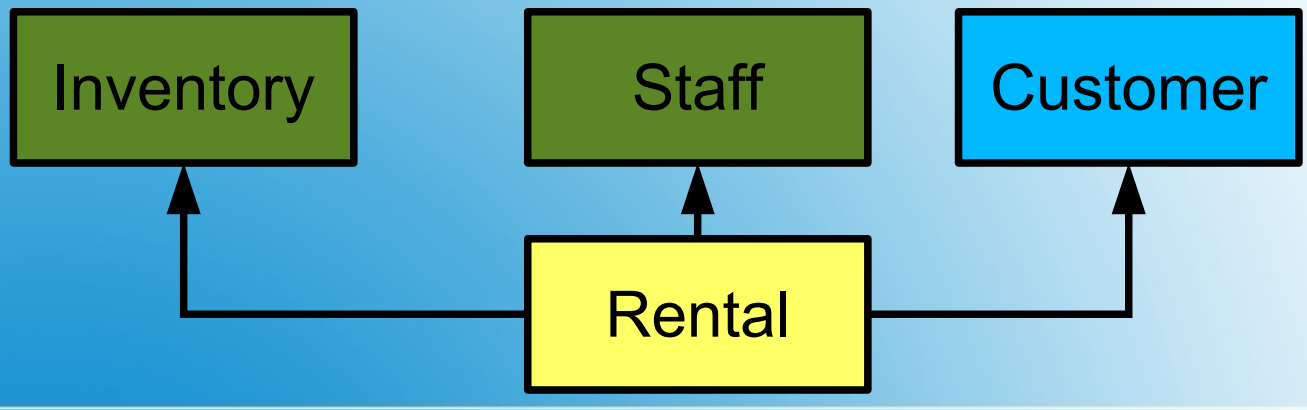
Target Star Schema

- Select Business Process
  - Sales, Purchase, Storage, ...
- Define Facts and Key Metrics
  - Facts: Key Event in Business Process
  - Metrics (Fact Attributes): Count or Amount
- Choose Dimensions and Hierarchies
  - What? When? Where?
  - Who? Why?

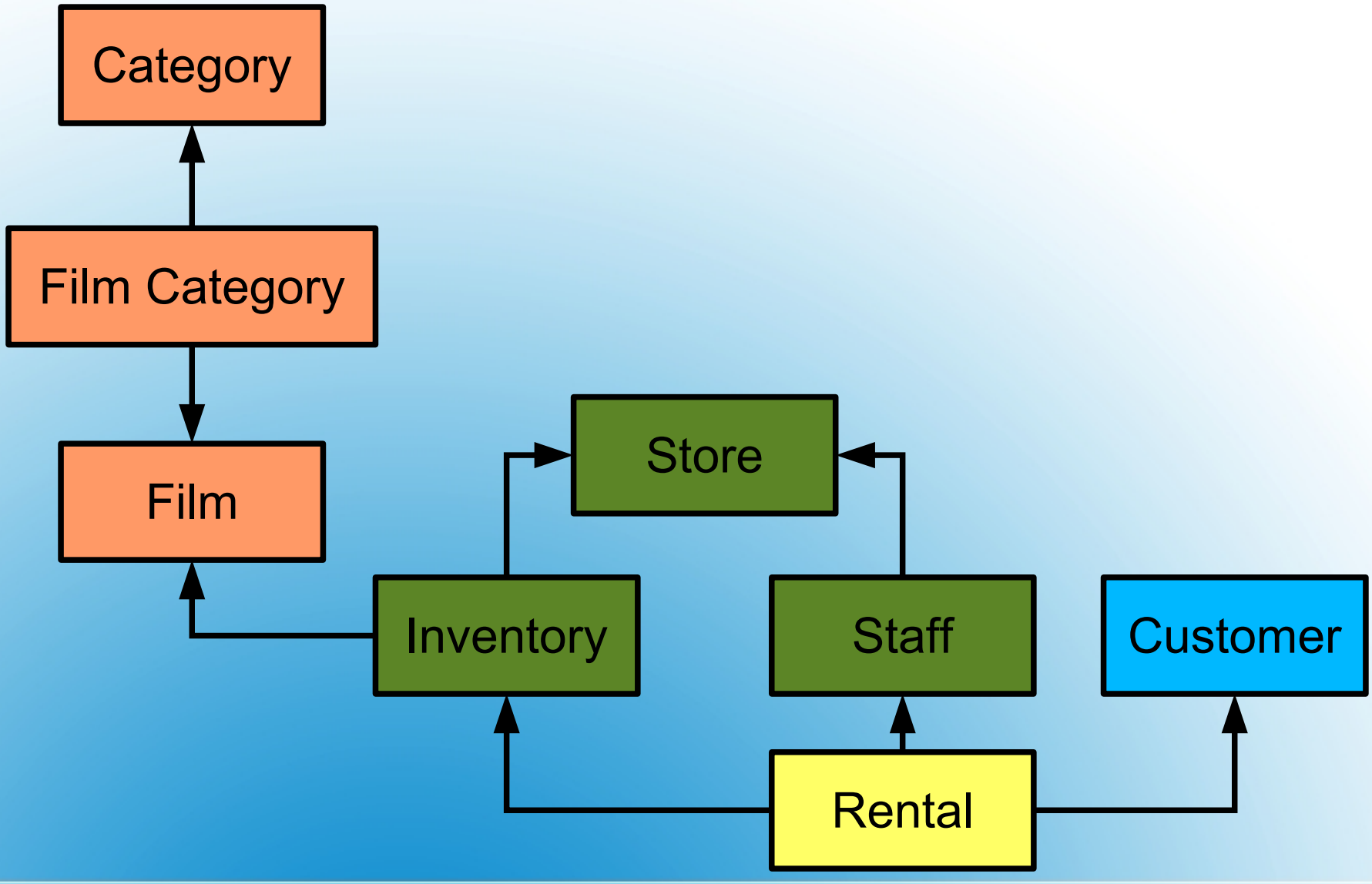
Dimensional Design

Example Business Process: Rentals

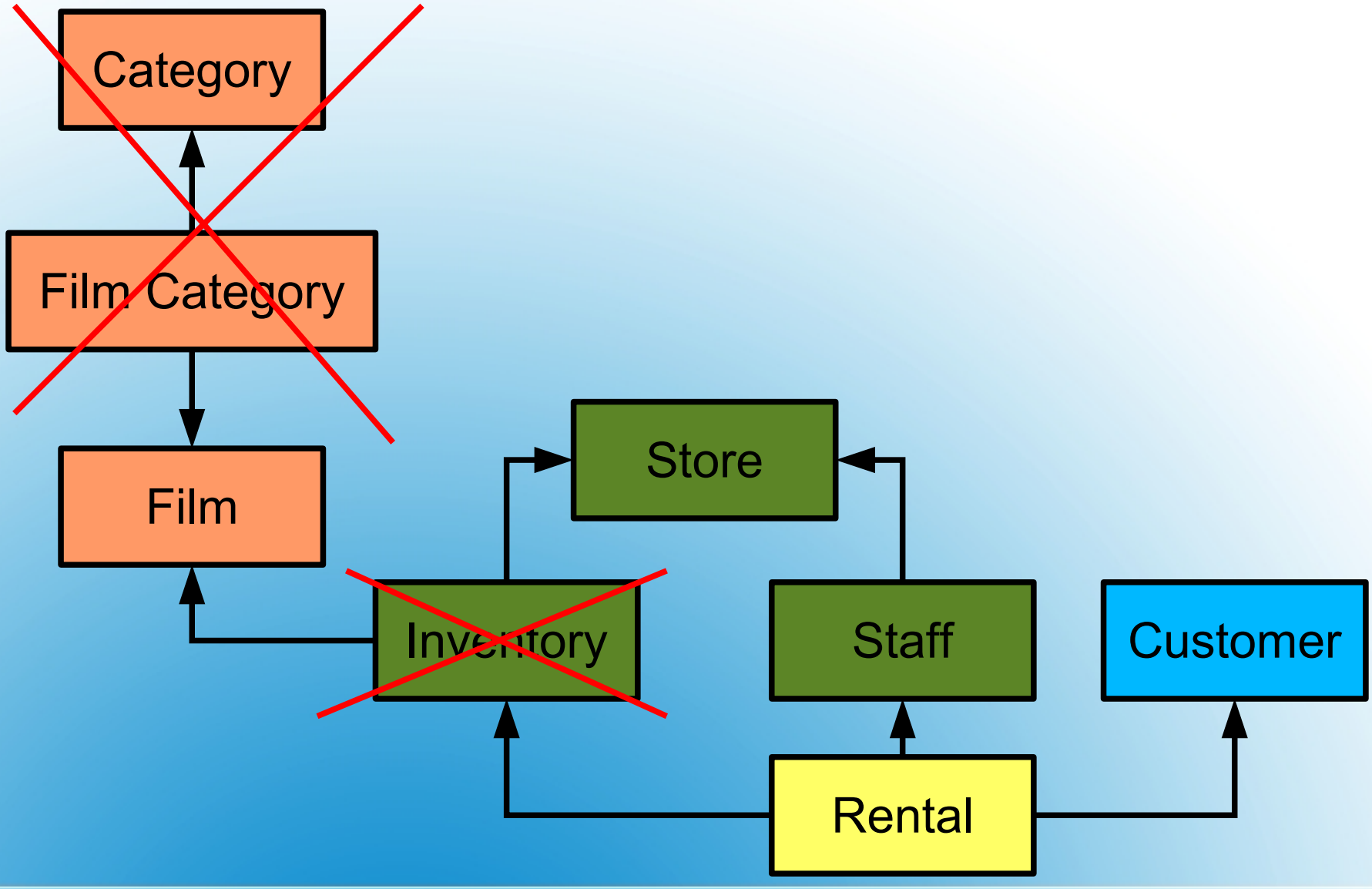
- Select Business Process
  - Rentals
- Identify Facts
  - Count (number of rentals)
  - Rental Duration
- Choose Dimensions
  - What: Films
  - When: Rental, Return
  - Who: Customer, Staff
  - Where: Store



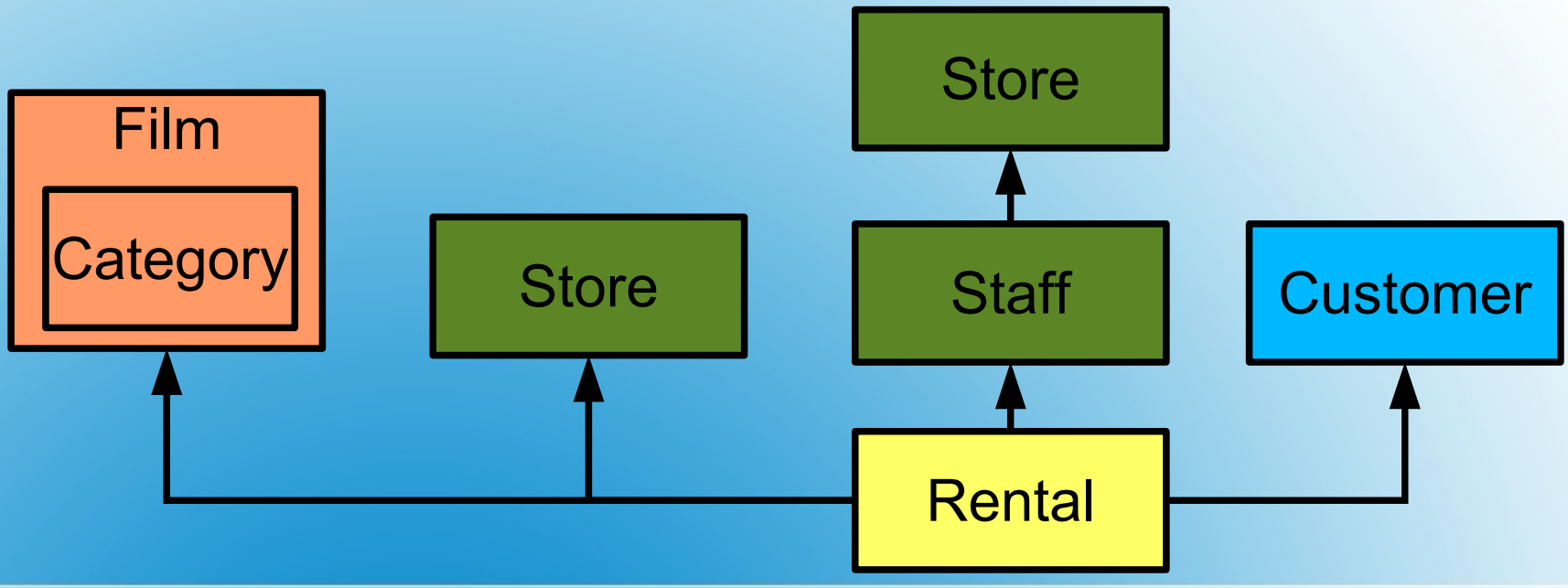
A star is born: Rentals 3NF



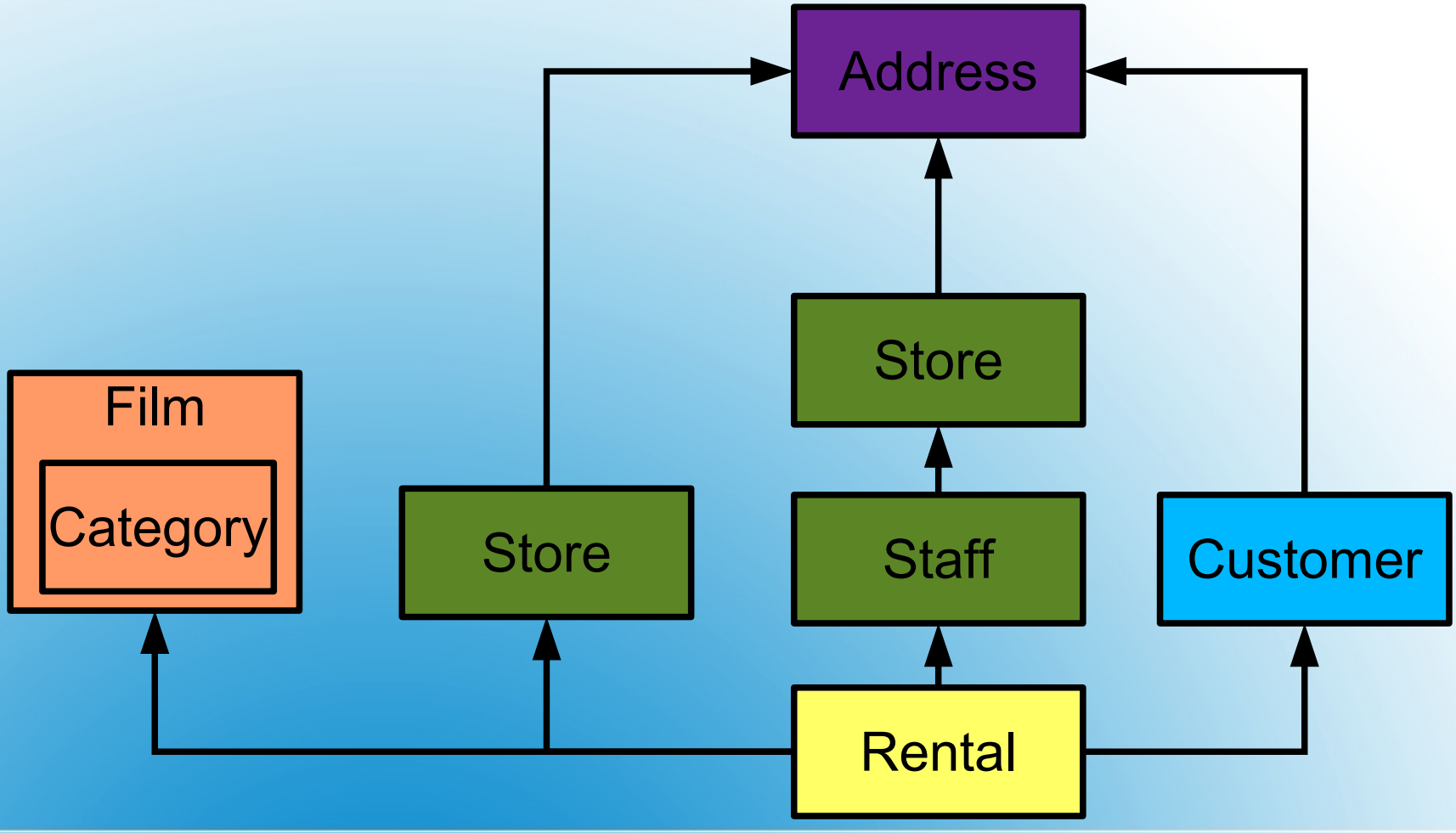
A star is born: Rentals 3NF



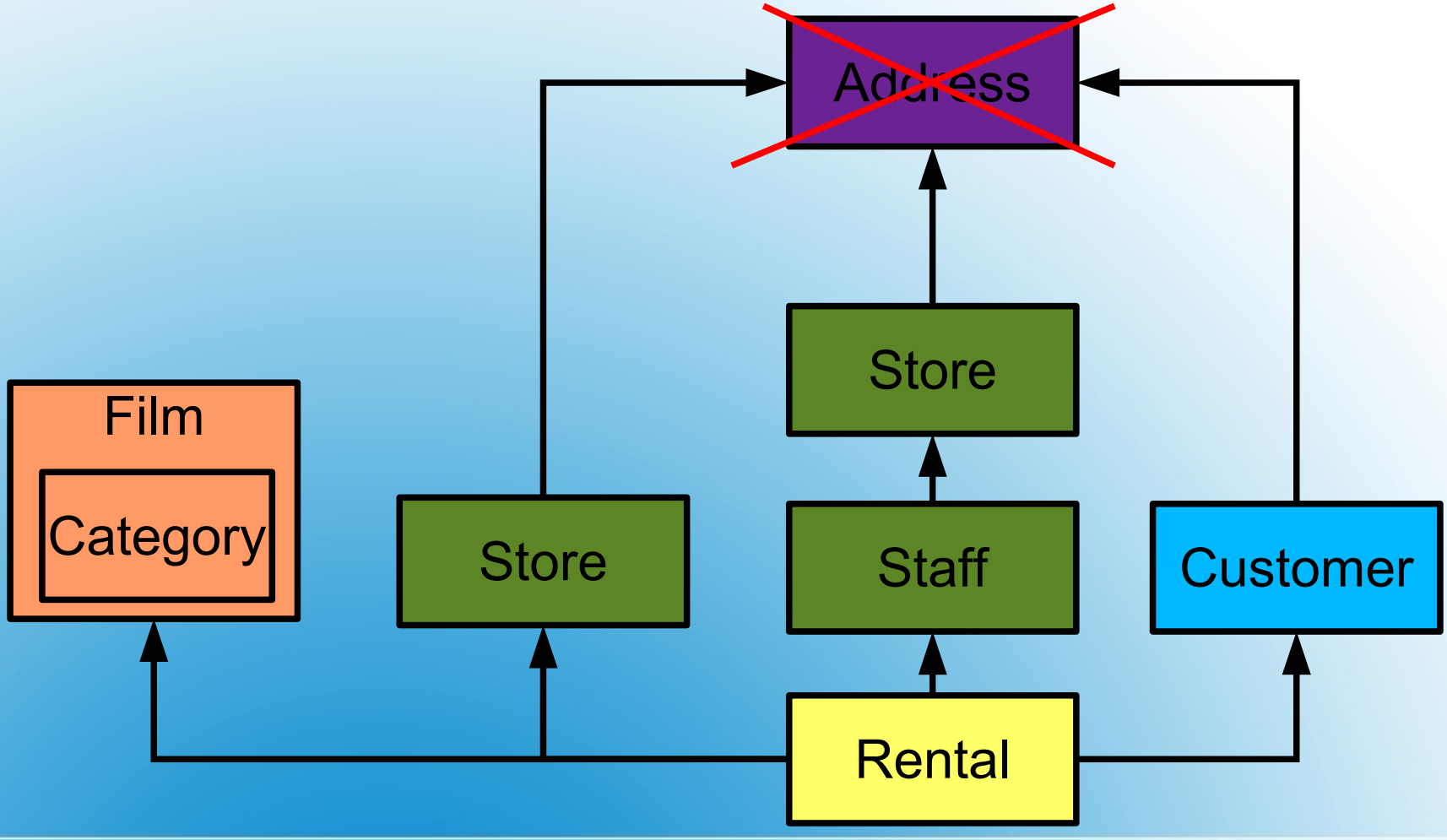
A star is born: Denormalize



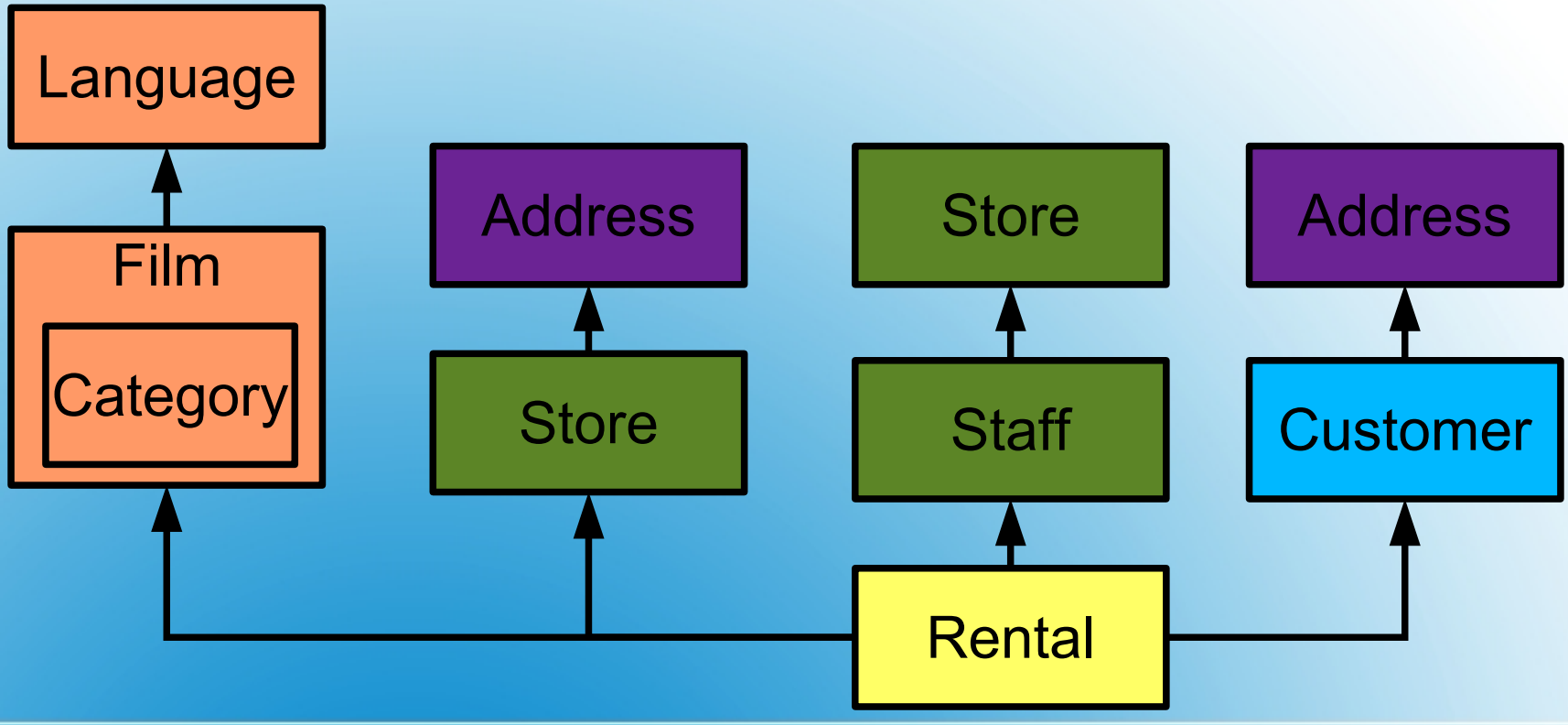
A star is born: Denormalize



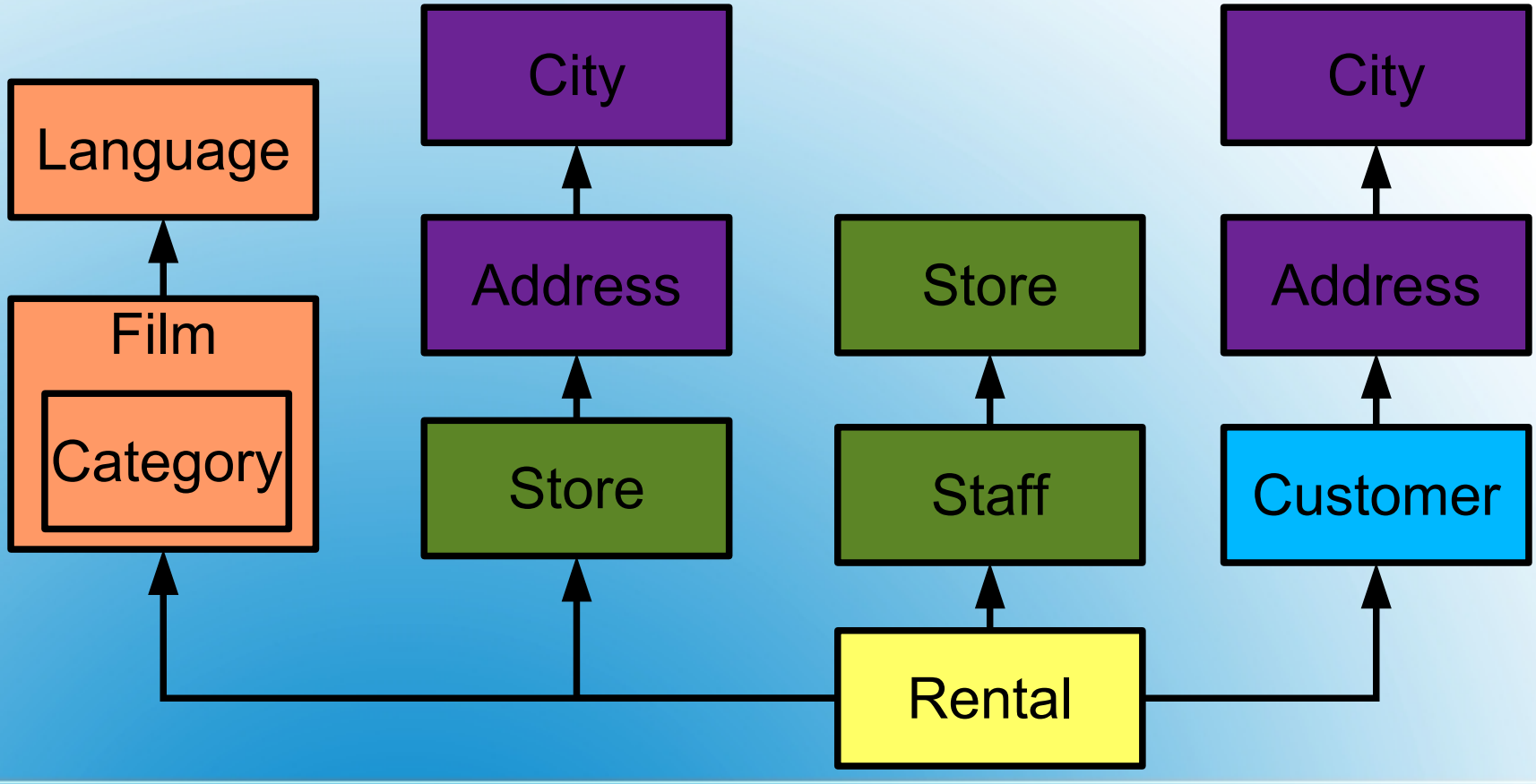
A star is born



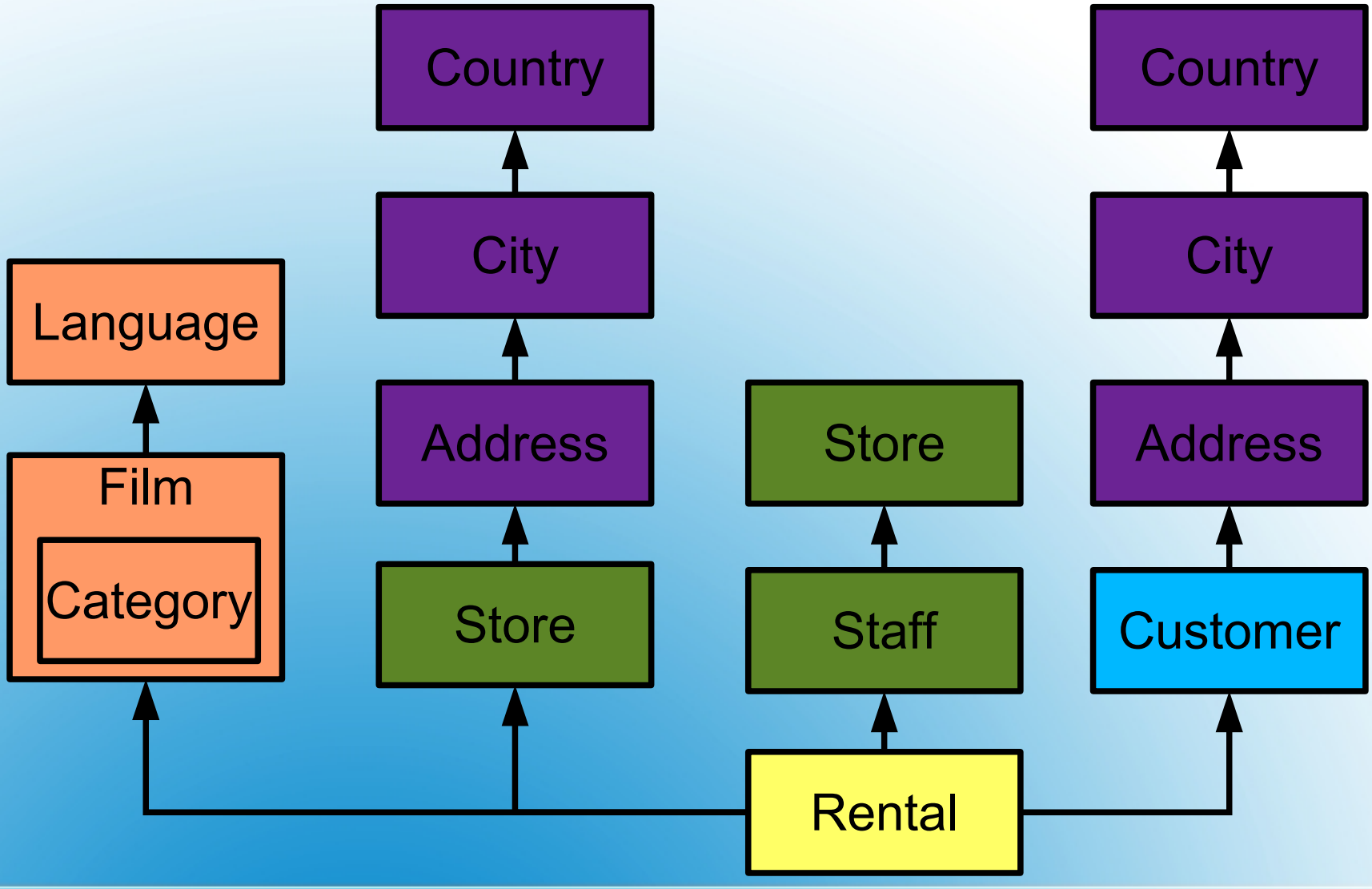
A star is born: Denormalize



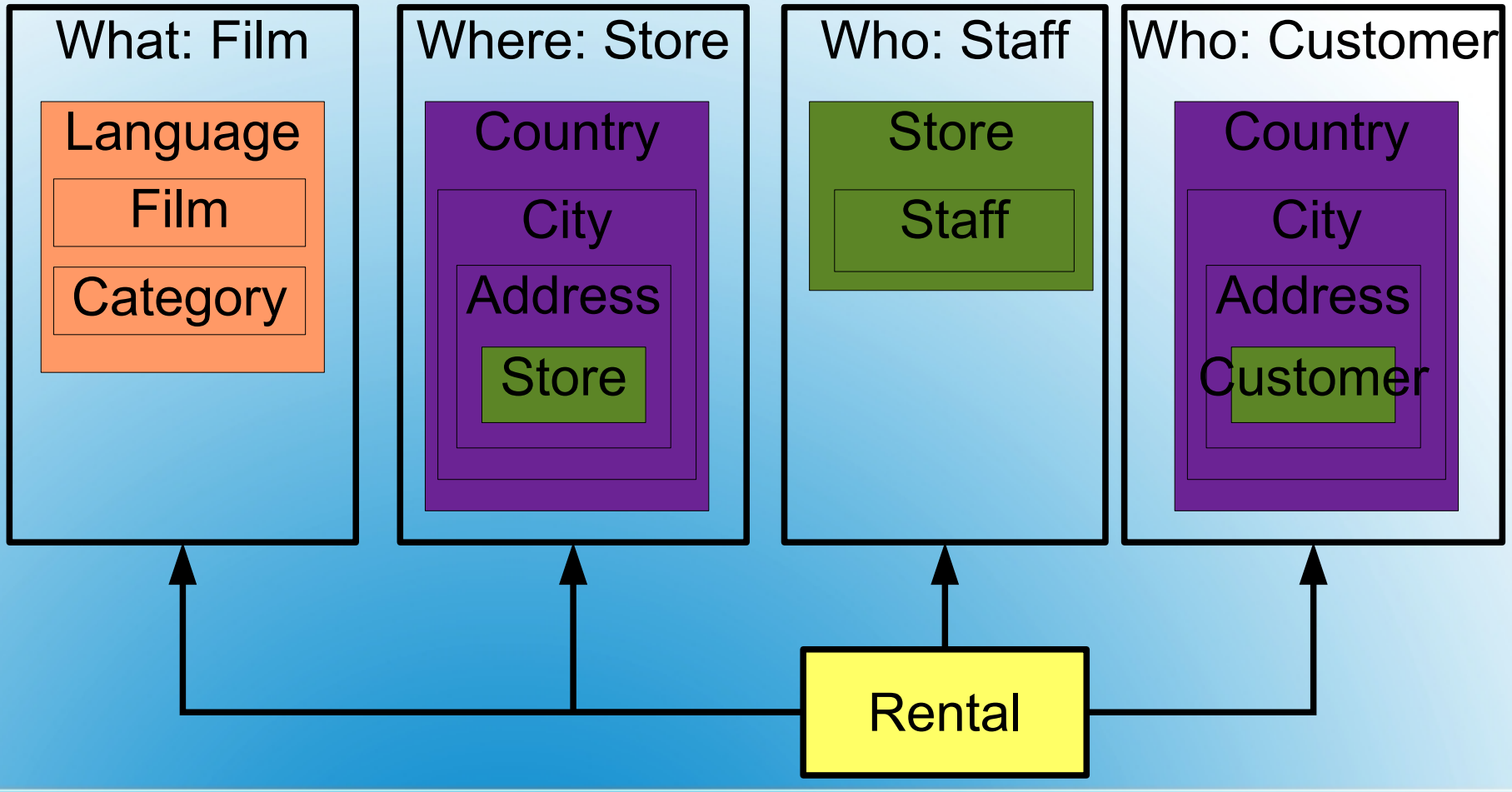
A star is born: Denormalize



A star is born: Denormalize



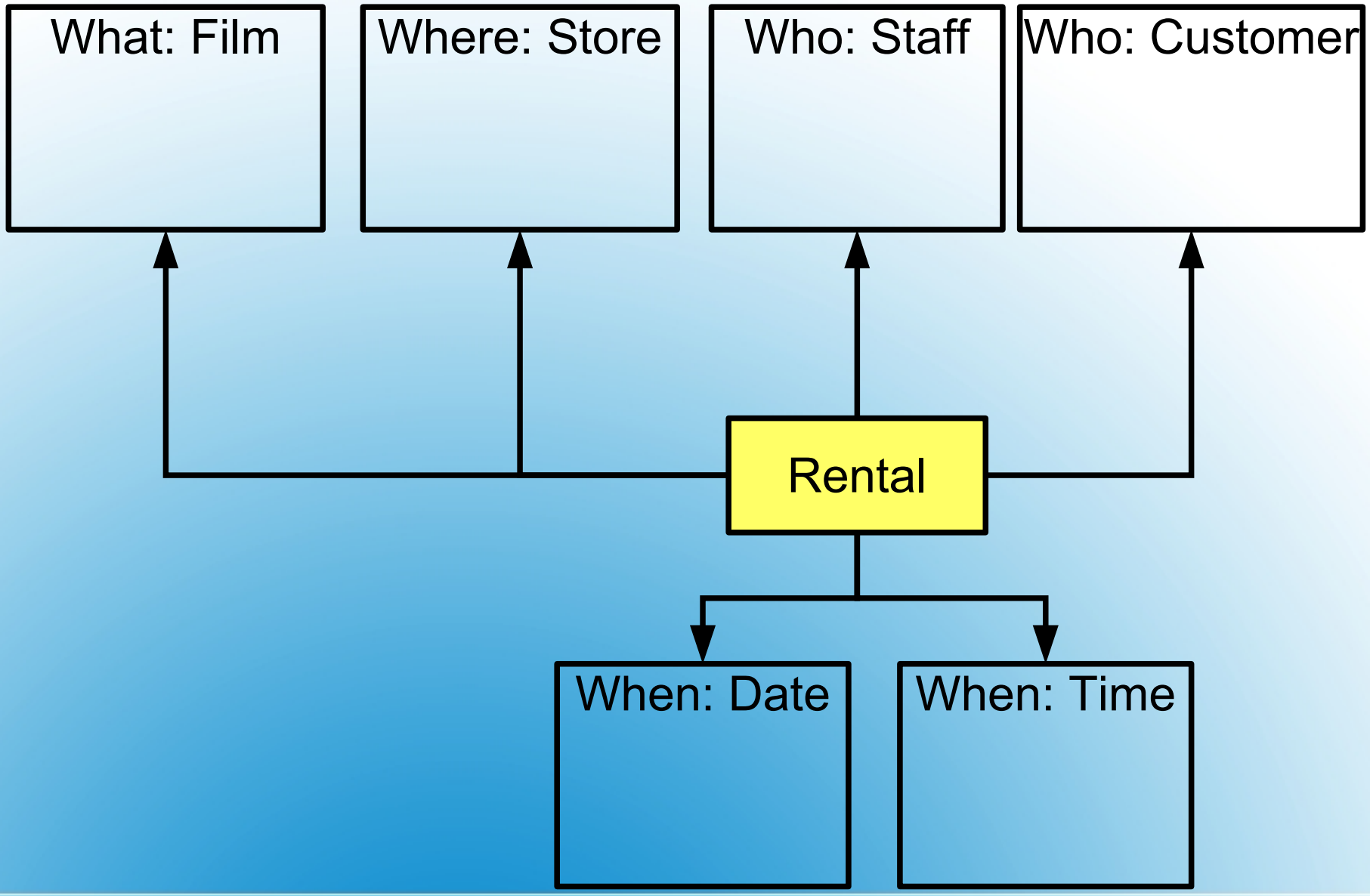
A star is born: Rental Snowflake



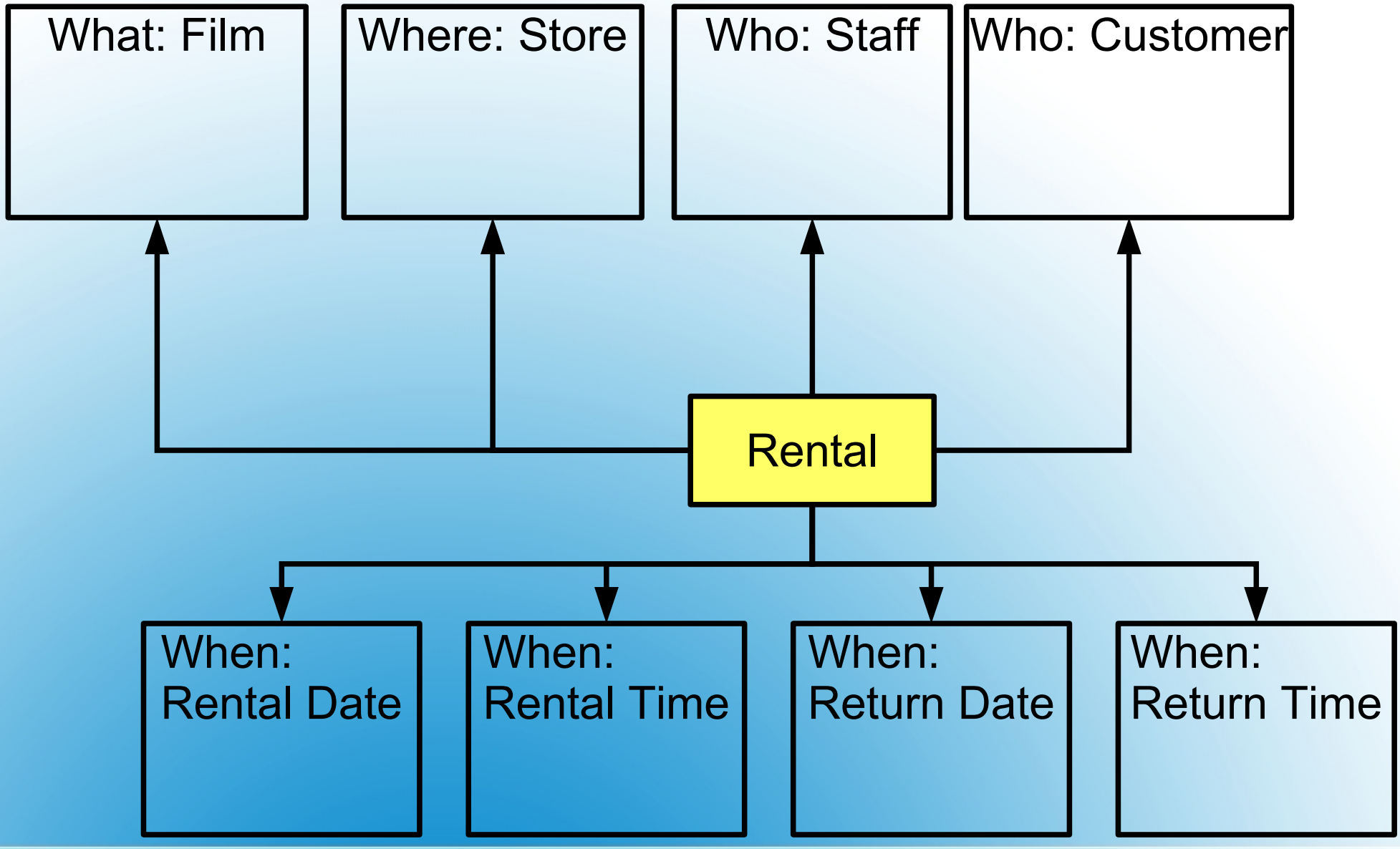
A star is born: Rental Star Schema

- Something is missing....
  - Who ? (Customer, Staff)
  - What ? (Film)
  - Where ? (Store)
  - .... ?

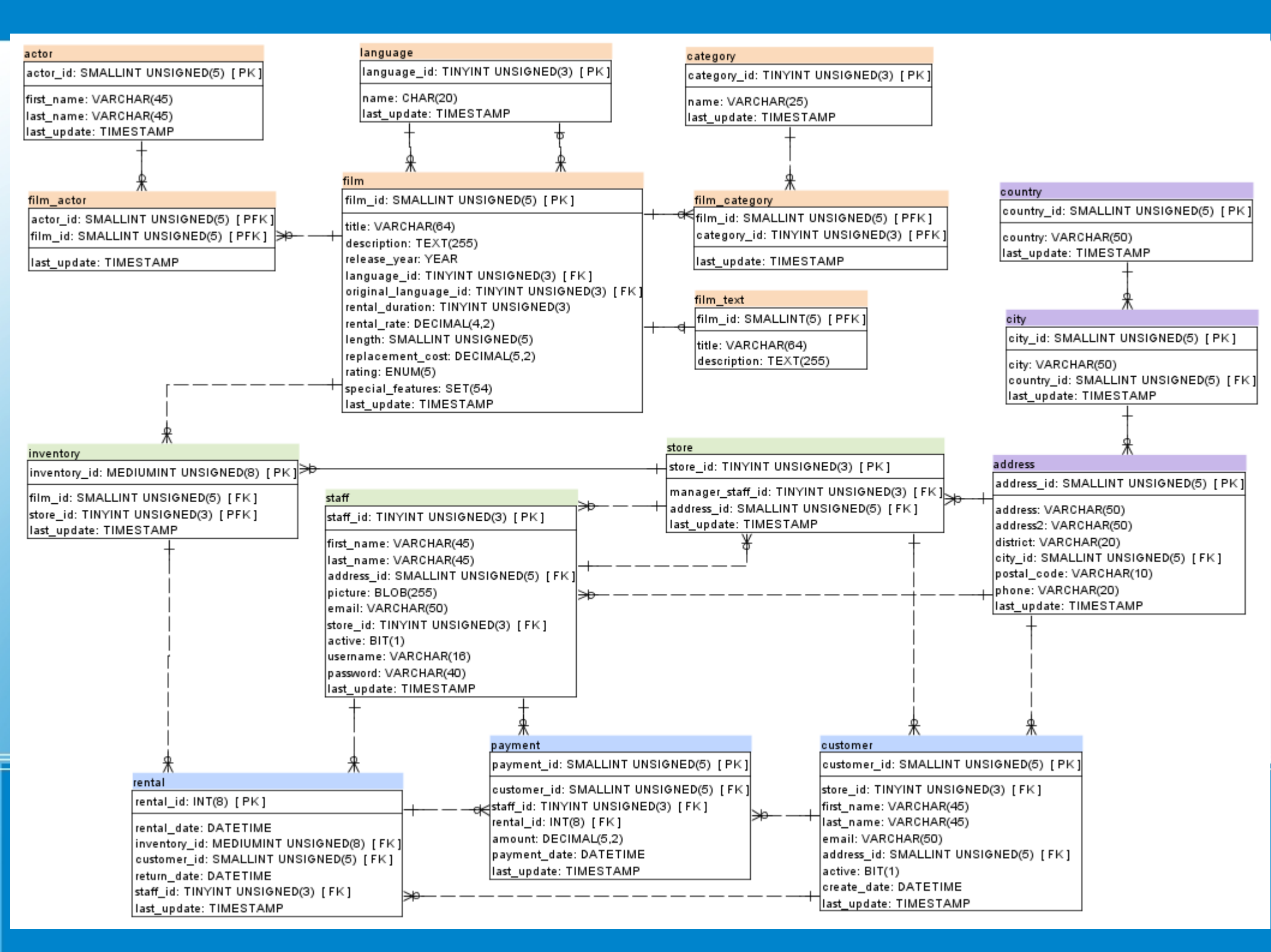
Dimensional Design

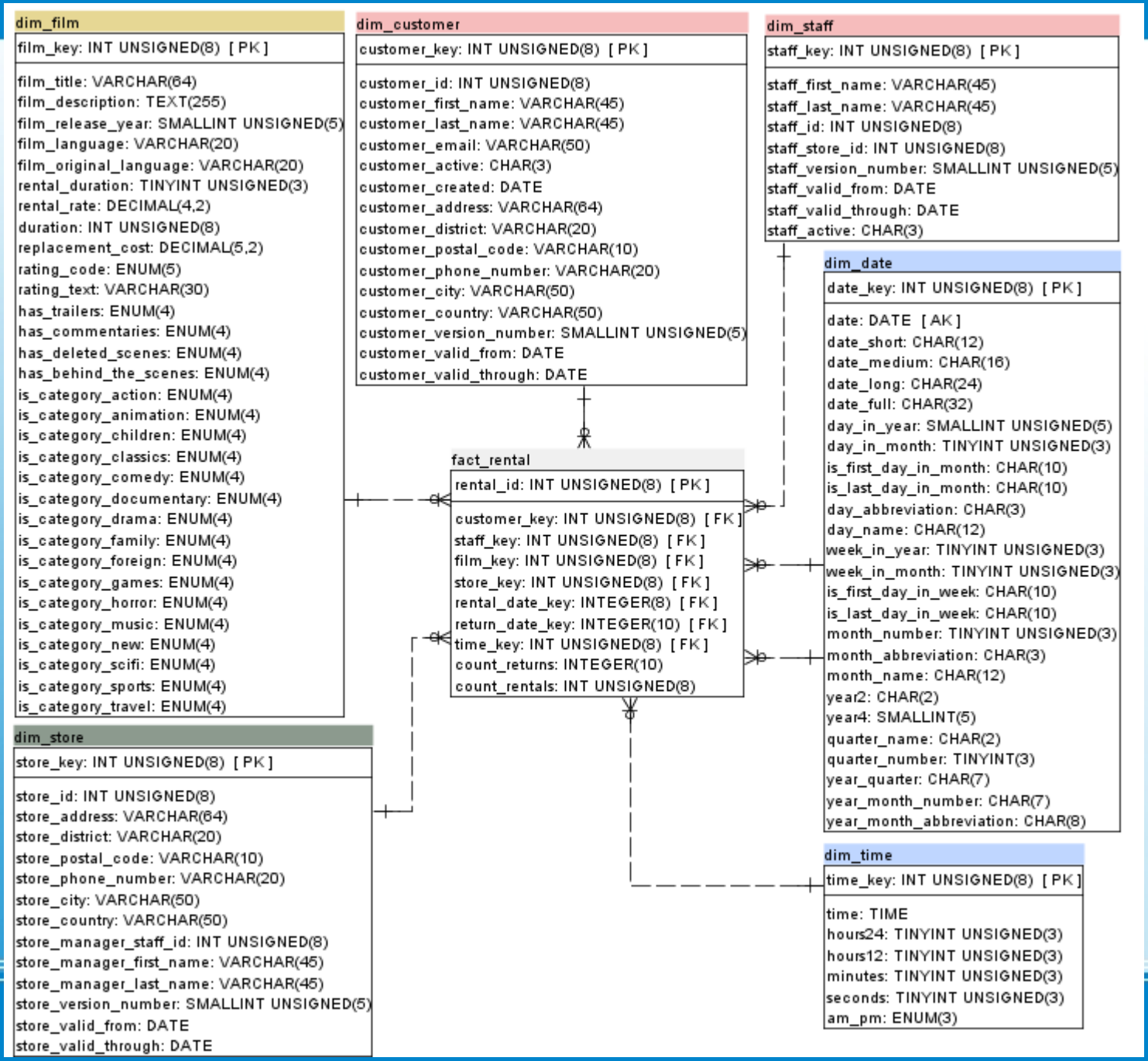


A star is born:
Rental Date and Time



Role Playing: Date/Time for both Rentals and Returns


actor
actor_id: SMALLINT UNSIGNED(5) [PK]
first_name: VARCHAR(45)
last_name: VARCHAR(45)
last_update: TIMESTAMP
language
language_id: TINYINT UNSIGNED(3) [PK]
name: CHAR(20)
last_update: TIMESTAMP
category
category_id: TINYINT UNSIGNED(3) [PK]
name: VARCHAR(25)
last_update: TIMESTAMP
film_actor
actor_id: SMALLINT UNSIGNED(5) [PFK]
film_id: SMALLINT UNSIGNED(5) [PFK]
last_update: TIMESTAMP
film
film_id: SMALLINT UNSIGNED(5) [PK]
title: VARCHAR(64)
description: TEXT(255)
release_year: YEAR
language_id: TINYINT UNSIGNED(3) [FK]
original_language_id: TINYINT UNSIGNED(3) [FK]
rental_duration: TINYINT UNSIGNED(3)
rental_rate: DECIMAL(4,2)
length: SMALLINT UNSIGNED(5)
replacement_cost: DECIMAL(5,2)
rating: ENUM(5)
special_features: SET(54)
last_update: TIMESTAMP
film_category
film_id: SMALLINT UNSIGNED(5) [PFK]
category_id: TINYINT UNSIGNED(3) [PFK]
last_update: TIMESTAMP
country
country_id: SMALLINT UNSIGNED(5) [PK]
country: VARCHAR(50)
last_update: TIMESTAMP
film_text
film_id: SMALLINT(5) [PFK]
title: VARCHAR(64)
description: TEXT(255)
city
city_id: SMALLINT UNSIGNED(5) [PK]
city: VARCHAR(50)
country_id: SMALLINT UNSIGNED(5) [FK]
last_update: TIMESTAMP
inventory
inventory_id: MEDIUMINT UNSIGNED(8) [PK]
film_id: SMALLINT UNSIGNED(5) [FK]
store_id: TINYINT UNSIGNED(3) [PFK]
last_update: TIMESTAMP
store
store_id: TINYINT UNSIGNED(3) [PK]
manager_staff_id: TINYINT UNSIGNED(3) [FK]
address_id: SMALLINT UNSIGNED(5) [FK]
last_update: TIMESTAMP
address
address_id: SMALLINT UNSIGNED(5) [PK]
address: VARCHAR(50)
address2: VARCHAR(50)
district: VARCHAR(20)
city_id: SMALLINT UNSIGNED(5) [FK]
postal_code: VARCHAR(10)
phone: VARCHAR(20)
last_update: TIMESTAMP
staff
staff_id: TINYINT UNSIGNED(3) [PK]
first_name: VARCHAR(45)
last_name: VARCHAR(45)
address_id: SMALLINT UNSIGNED(5) [FK]
picture: BLOB(255)
email: VARCHAR(50)
store_id: TINYINT UNSIGNED(3) [FK]
active: BIT(1)
username: VARCHAR(16)
password: VARCHAR(40)
last_update: TIMESTAMP
payment
payment_id: SMALLINT UNSIGNED(5) [PK]
customer_id: SMALLINT UNSIGNED(5) [FK]
staff_id: TINYINT UNSIGNED(3) [FK]
rental_id: INT(8) [FK]
amount: DECIMAL(5,2)
payment_date: DATETIME
last_update: TIMESTAMP
customer
customer_id: SMALLINT UNSIGNED(5) [PK]
store_id: TINYINT UNSIGNED(3) [FK]
first_name: VARCHAR(45)
last_name: VARCHAR(45)
email: VARCHAR(50)
address_id: SMALLINT UNSIGNED(5) [FK]
active: BIT(1)
create_date: DATETIME
last_update: TIMESTAMP
rental
rental_id: INT(8) [PK]
rental_date: DATETIME
inventory_id: MEDIUMINT UNSIGNED(8) [FK]
customer_id: SMALLINT UNSIGNED(5) [FK]
return_date: DATETIME
staff_id: TINYINT UNSIGNED(3) [FK]
last_update: TIMESTAMP

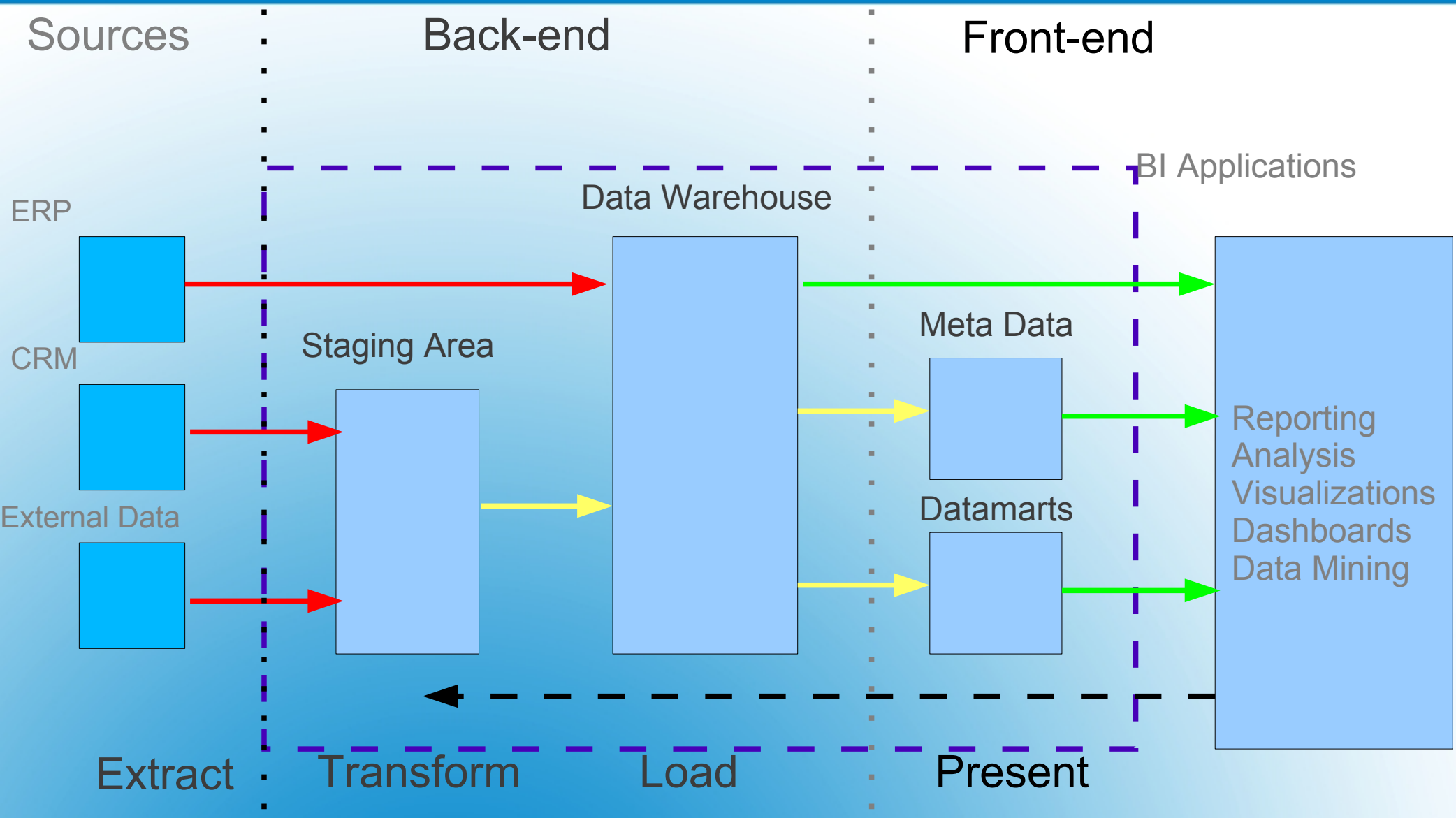


Rental Star Schema

Part IV:

Filling the Data Warehouse

Starring Sakila



High Level Data Warehouse Architecture

Planning the ETL Process

- Physical Design
- Source to Target Mapping
  - Define how data in the data warehouse is derived from data in the source system(s)
  - Specification for designing the ETL process
- Column-level mapping
  - Source system, schema, table, column, data type
  - Target dimension/fact, column, defaults
  - Transformation rules, cleansing, lookup, calculation

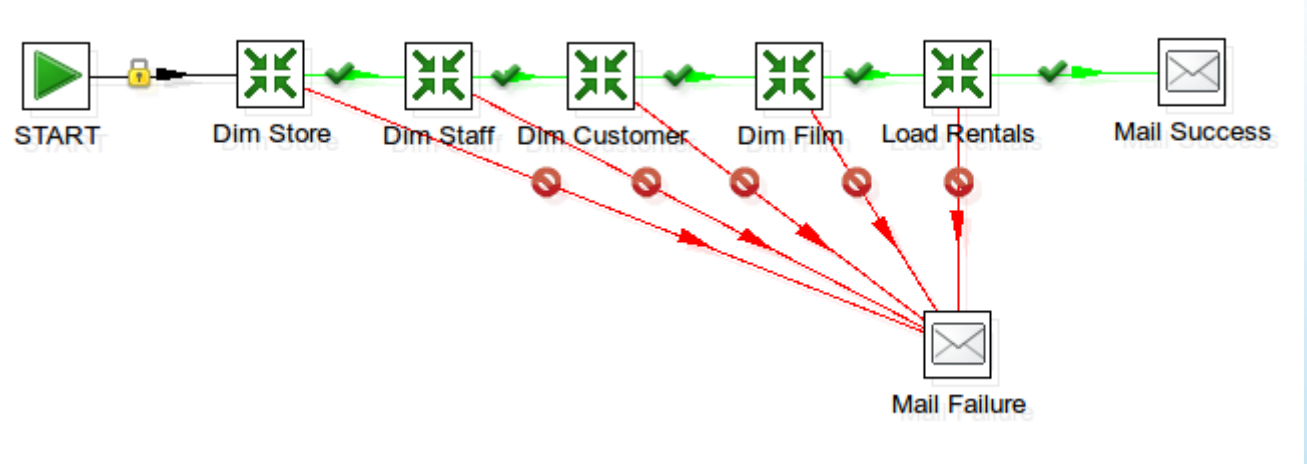
- Staging?
- Changed Data Capture / Extraction
- Denormalization
- Derived data / Enrichment
- Cleansing / Conforming
- History policy (dimensions)
- Granularity
- Dimension Lookup (facts)

Designing the ETL Process

- Flow ETL Engine
- Transformations
  - Data flow and processing
- Jobs
  - Workflow of ETL tasks
- Tools
  - Spoon
  - Kitchen
  - Pan

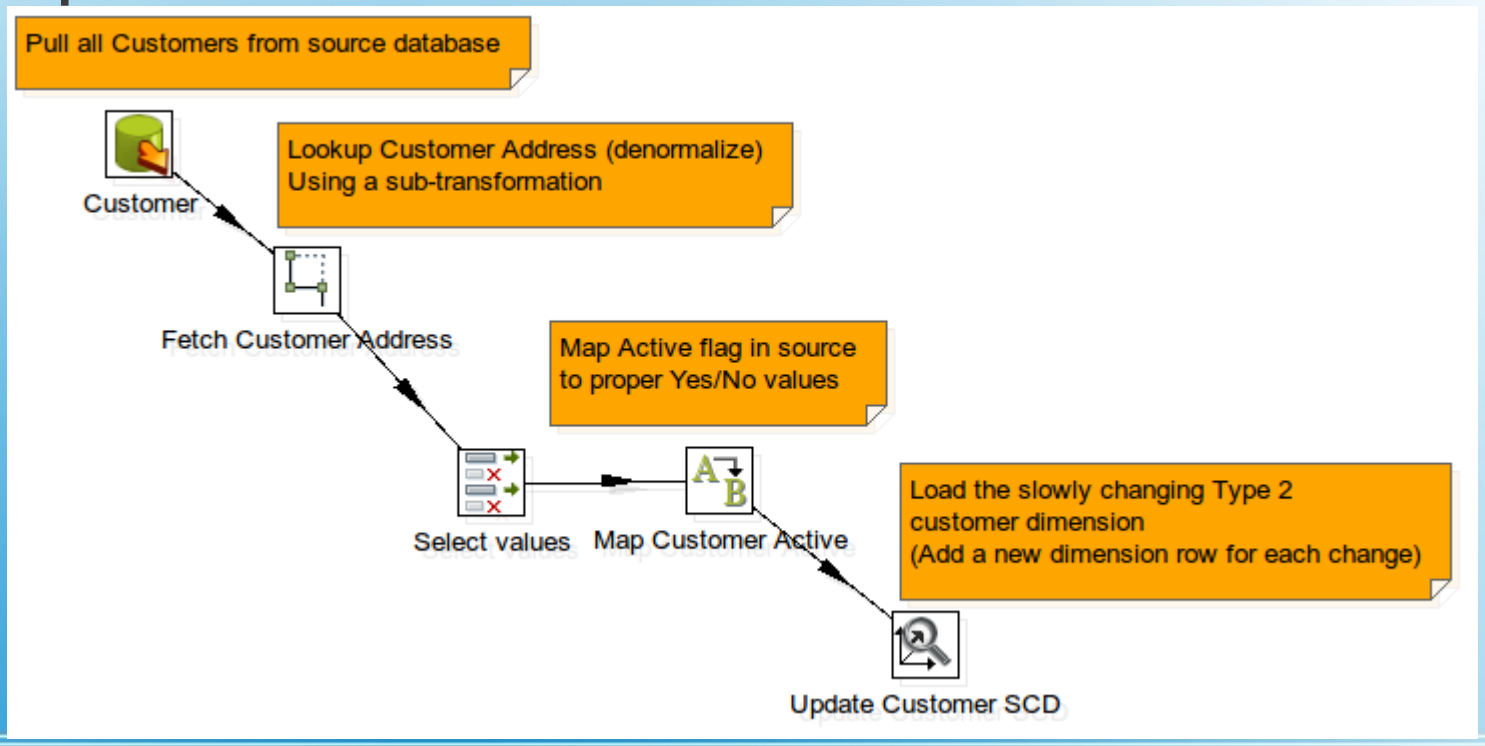
Designing ETL with Kettle

- Load Dimension Tables
- Load Fact table

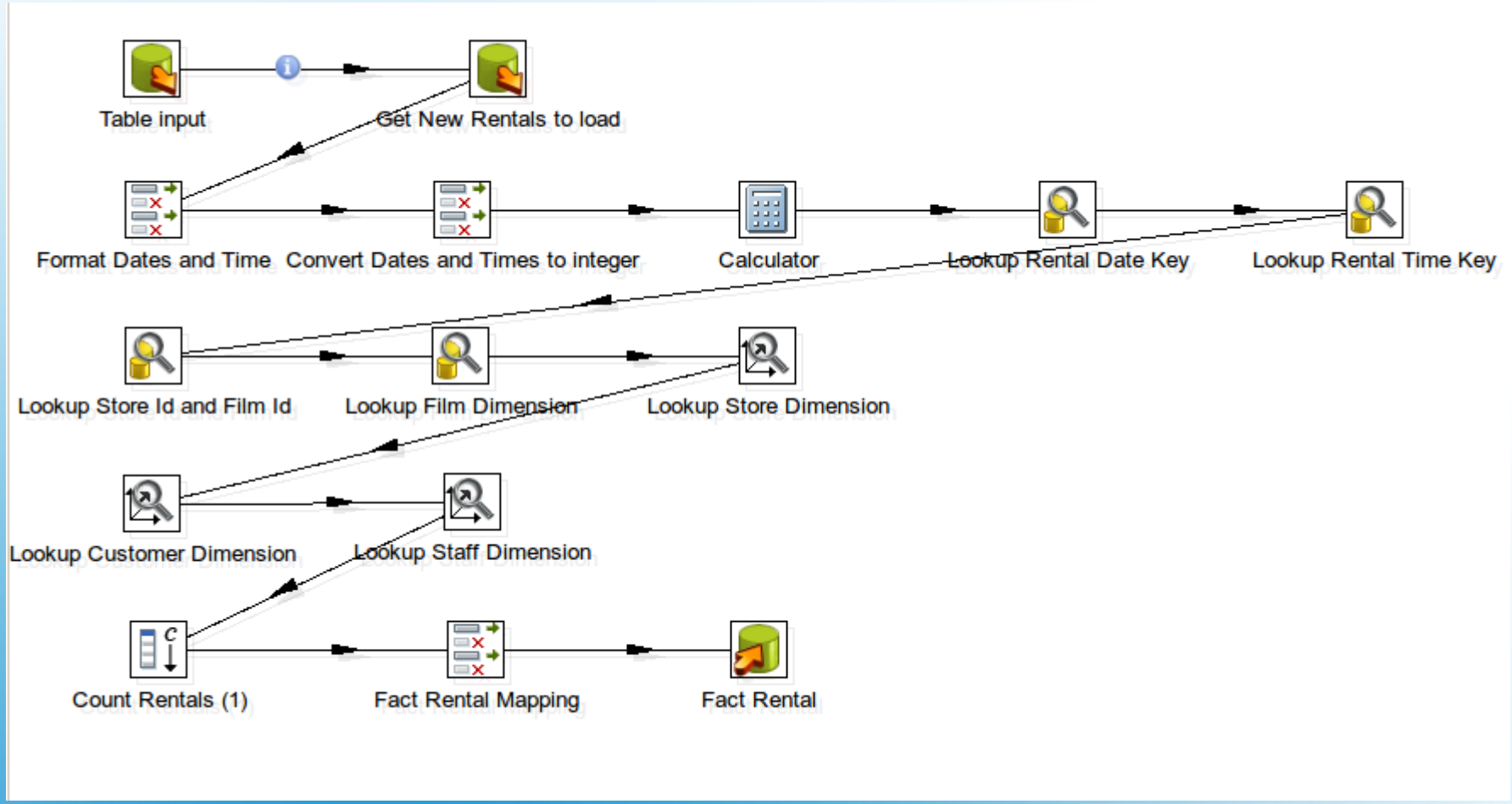


Loading a Fact Table

- Get Customers source data
- Lookup Address (Denormalize)
- Update Dimension



Loading a Dimension Table



Loading a Fact Table

Part V:

Presenting the Data: BI Applications

Starring Sakila

months, years:

Should we become an appliance vendor instead of delivering software solutions

weeks, months:

In what region should we open a new store?

days, weeks:

Who's available for tomorrow's shift

Data mining

OLAP/Analysis

Reporting

Business Intelligence Scope

Reporting

Reporting

- Mostly Operational
- Lists and Grouping
- Typically standardized
- Typically no or limited interactivity
  - Subreporting

months, years:

Should we become an appliance vendor instead of delivering software solutions

weeks, months:

In what region should we open a new store?

days, weeks:

Who's available for tomorrow's shift

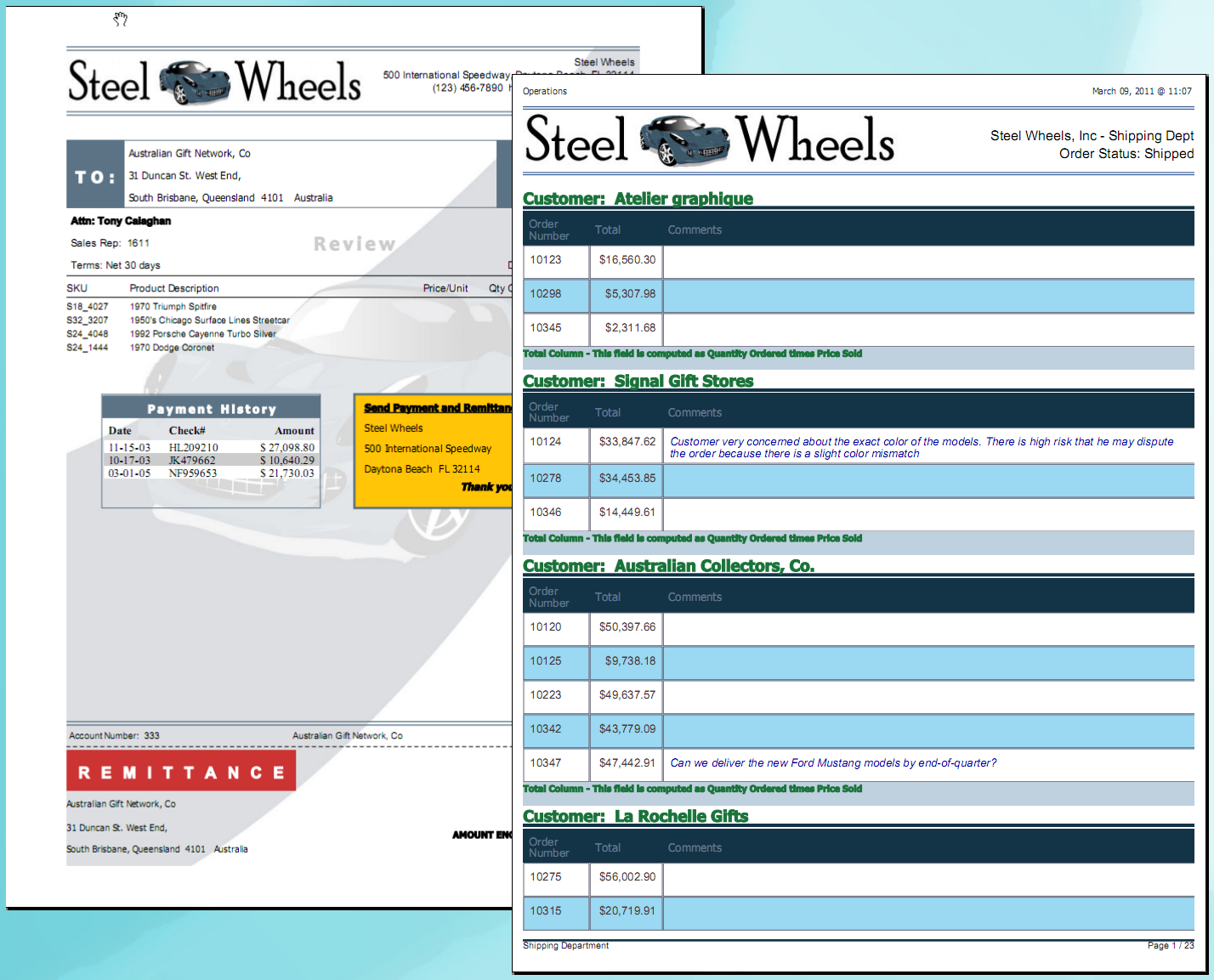
Data mining

OLAP/Analysis

Reporting

Scope of Reporting

Reporting



Steel Wheels

500 International Speedway
(123) 456-7890

TO: Australian Gift Network, Co
31 Duncan St. West End,
South Brisbane, Queensland 4101 Australia

Attn: Tony Calaghan

Sales Rep: 1611

Terms: Net 30 days

Review

SKU	Product Description	Price/Unit	Qty
S18_4027	1970 Triumph Spitfire		
S32_3207	1950's Chicago Surface Lines Streetcar		
S24_4048	1992 Porsche Cayenne Turbo Silver		
S24_1444	1970 Dodge Coronet		

Payment History

Date	Check#	Amount
11-15-03	HL209210	$ 27,098.80
10-17-03	JK479662	$ 10,640.29
03-01-05	NF959653	$ 21,730.03

Send Payment and Remittan
Steel Wheels
500 International Speedway
Daytona Beach FL 32114
Thank yo

Account Number: 333 Australian Gift Network, Co

REMITTANCE

Australian Gift Network, Co
31 Duncan St. West End,
South Brisbane, Queensland 4101 Australia

AMOUNT EN

Operations March 09, 2011 @ 11:07

Steel Wheels

Steel Wheels, Inc - Shipping Dept
Order Status: Shipped

Customer: Atelier graphique

Order Number	Total	Comments
10123	$16,560.30	
10298	$5,307.98	
10345	$2,311.68	

Total Column - This field is computed as Quantity Ordered times Price Sold

Customer: Signal Gift Stores

Order Number	Total	Comments
10124	$33,847.62	*Customer very concerned about the exact color of the models. There is high risk that he may dispute the order because there is a slight color mismatch*
10278	$34,453.85	
10346	$14,449.61	

Total Column - This field is computed as Quantity Ordered times Price Sold

Customer: Australian Collectors, Co.

Order Number	Total	Comments
10120	$50,397.66	
10125	$9,738.18	
10223	$49,637.57	
10342	$43,779.09	
10347	$47,442.91	*Can we deliver the new Ford Mustang models by end-of-quarter?*

Total Column - This field is computed as Quantity Ordered times Price Sold

Customer: La Rochelle Gifts

Order Number	Total	Comments
10275	$56,002.90	
10315	$20,719.91	

Shipping Department Page 1 / 23

Analysis

Analysis

- Tactical, Strategic
- OLAP
  - Online Analytical Processing
- Pivot tables
- Typically Interactive
  - Slice and Dice
  - Drilldown
- Typically Ad-hoc

months, years:

Should we become an appliance vendor instead of delivering software solutions

weeks, months:

In what region should we open a new store?

days, weeks:

Who's available for tomorrow's shift

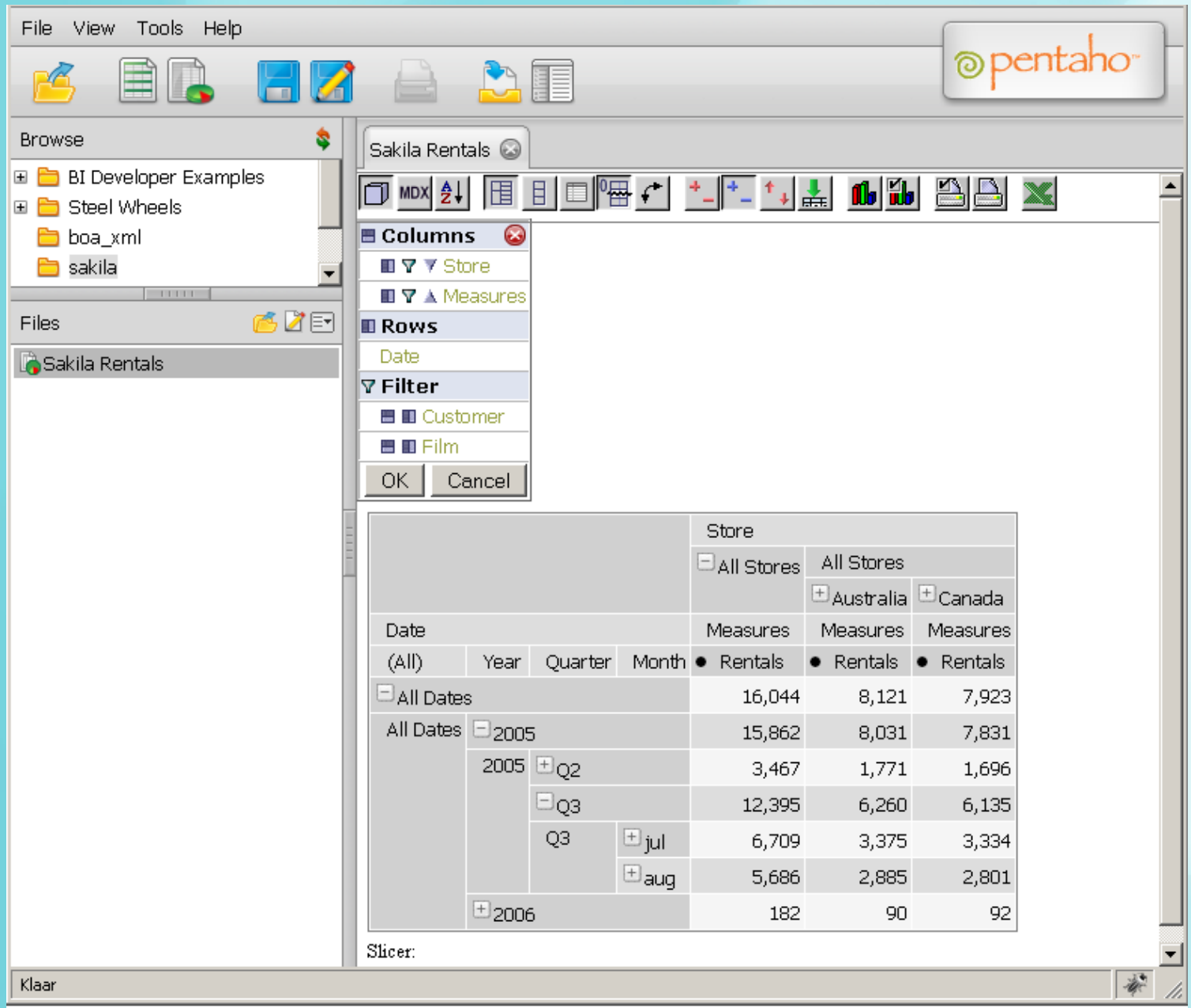
Data mining

OLAP/Analysis

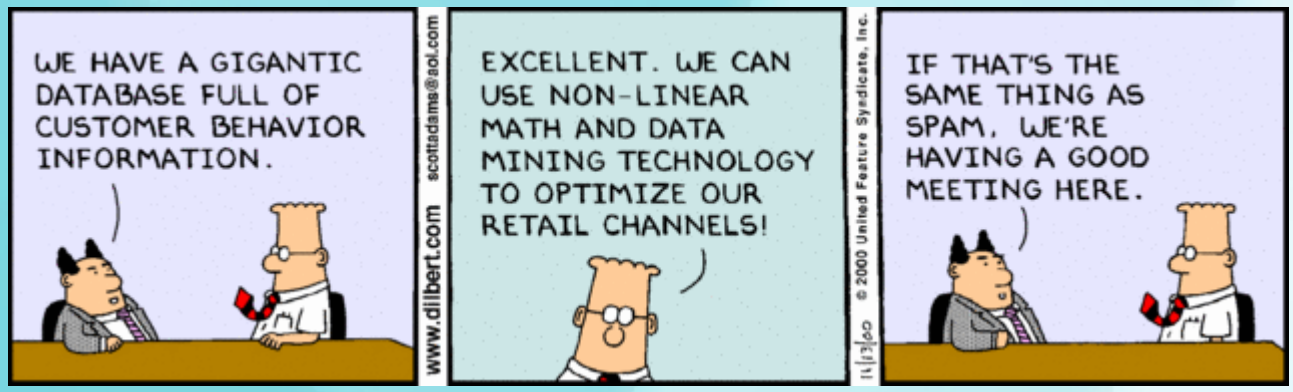
Reporting

Scope of OLAP & Analysis

Analysis Interactive Pivot table



Data Mining


WE HAVE A GIGANTIC DATABASE FULL OF CUSTOMER BEHAVIOR INFORMATION.
EXCELLENT. WE CAN USE NON-LINEAR MATH AND DATA MINING TECHNOLOGY TO OPTIMIZE OUR RETAIL CHANNELS!
IF THAT'S THE SAME THING AS SPAM, WE'RE HAVING A GOOD MEETING HERE.
www.dilbert.com scottadams@aol.com
© 2000 United Feature Syndicate, Inc.

Data Mining

- Strategic, Tactical
- Discover hidden patterns in data
- Machine learning
- Statistic analysis
- Typically not interactive, long running
- Expert matter
- Not readily consumable by end-users
  - Characteristics of back-end processing

months, years:

Should we become an appliance vendor instead of delivering software solutions

Data mining

weeks, months:

In what region should we open a new store?

OLAP/Analysis

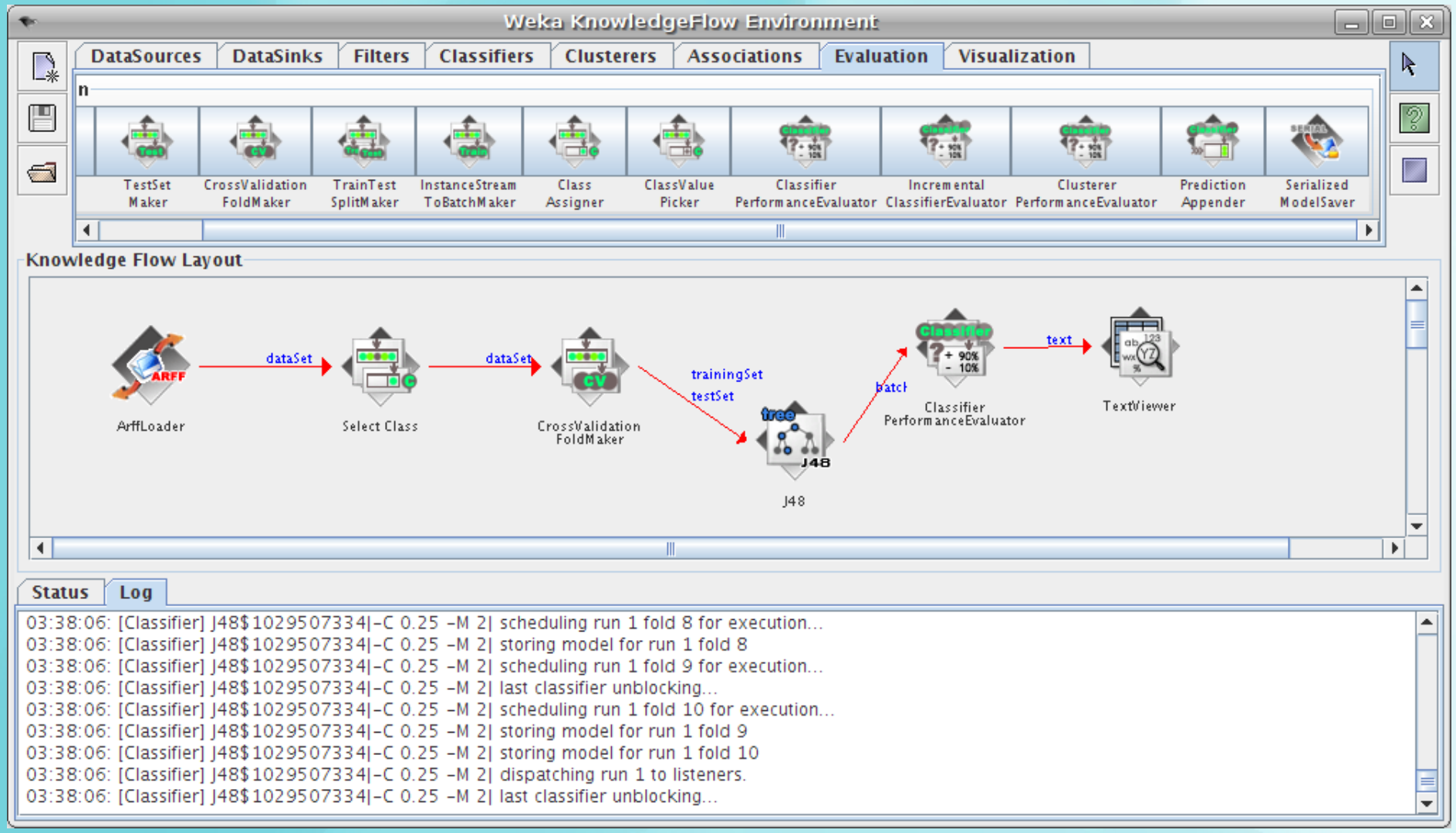
days, weeks:

Who's available for tomorrow's shift

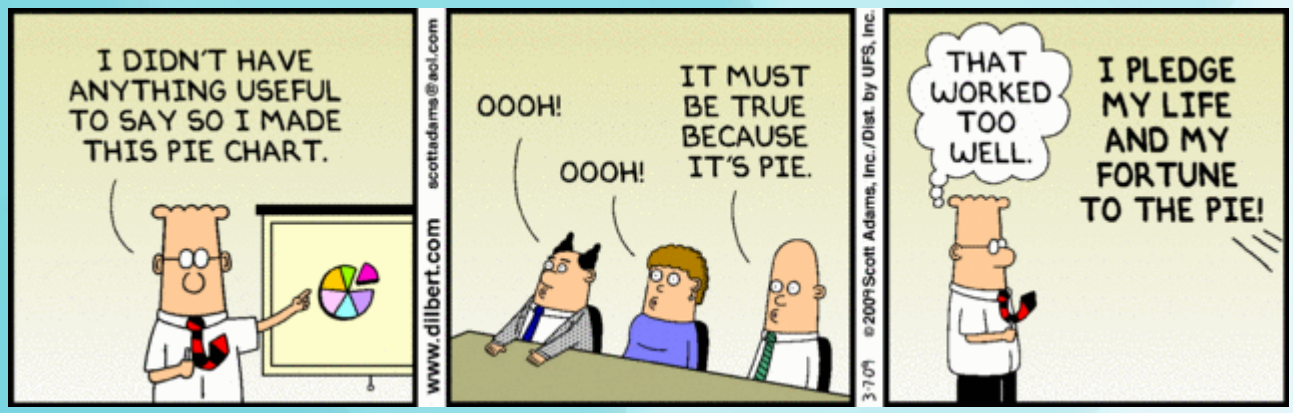
Reporting

Scope of Data Mining

Data Mining


Weka KnowledgeFlow Environment
DataSources
DataSinks
Filters
Classifiers
Clusterers
Associations
Evaluation
Visualization
TestSet Maker
CrossValidation FoldMaker
TrainTest SplitMaker
InstanceStream ToBatchMaker
Class Assigner
ClassValue Picker
Classifier PerformanceEvaluator
Incremental ClassifierEvaluator
Clusterer PerformanceEvaluator
Prediction Appender
Serialized ModelSaver
Knowledge Flow Layout
ArffLoader
dataSet
Select Class
dataSet
CrossValidation FoldMaker
trainingSet
testSet
J48
batch
Classifier PerformanceEvaluator
text
TextViewer
Status
Log
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| scheduling run 1 fold 8 for execution...
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| storing model for run 1 fold 8
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| scheduling run 1 fold 9 for execution...
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| last classifier unblocking...
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| scheduling run 1 fold 10 for execution...
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| storing model for run 1 fold 9
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| storing model for run 1 fold 10
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| dispatching run 1 to listeners.
03:38:06: [Classifier] J48$1029507334|-C 0.25 -M 2| last classifier unblocking...

Charts and Graphs



Charts and Graphs

- Operational, Tactical, Strategic
- Summarize large dataset
- Not a separate class but a presentation
  - Data Visualization
- Standardized or ad-hoc
- Can be interactive
  - Drive a subreport
  - Drive drilldown

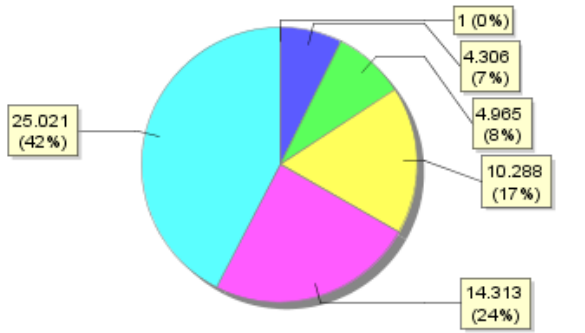
Dashboarding

Dashboarding

- Operational, Tactical, Strategic
- Not a separate class but a presentation
- Bundle:
  - key metrics for a particular role or perspective
  - different views on the same metrics
- Can contain reports, pivot tables, charts, graphs
- Typically interactive

Dashboard

Customers of Website: World Class Movies



Customers per Website

Top 100 Customer Locations for this Website