
Pivotal
A NEW PLATFORM FOR A NEW ERA

Data Modeling Considerations in GPDB

Pivotal® Greenplum Database

Agenda

- Introduction
- Models & Background
- Pivotal Approach
- Possible Optimizations

Pivotal

Introduction

- Greenplum database is not limited to specific models as some other database platforms are
- Greenplum supports features that are used in modeling:
  - Uniqueness
  - Constraints
  - Keys
- Greenplum does not support some features that are frequently used in modeling:
  - Foreign keys
  - Relational Integrity checks

Introduction (cont'd.)

- Just because Greenplum supports items like primary keys and constraints does not mean they should be used pervasively
- Also, just because your organization spent years modeling the existing warehouse does not mean it is functional for the **users** of the system
- Bill Inmon and Ralph Kimball practiced and developed their 'schemes' in a world before MPP. Their models and approaches reflect the technical time in which they were developed

Key Modeling Considerations

The following are key modeling considerations:

- Consider the business first
- Model for usability NOT to fit a model
- Do not force data into a model which does not fit the use
- Leave relational integrity checking to the source system
- Iterative and agile modeling is recommended over archaic and inflexible/prescriptive methods
- Legacy models will likely not be optimal*

* Just because you have worked for years on your optimal DW model does not mean it will work for the business or perform well

Pivotal

Historical Approaches

As mentioned previously, the models from Inmon and Kimball have been subscribed to as "the way to architect a data warehouse". These models:

- Star Schema
- Snowflake Schema

have been used to supplant the classic database pattern:

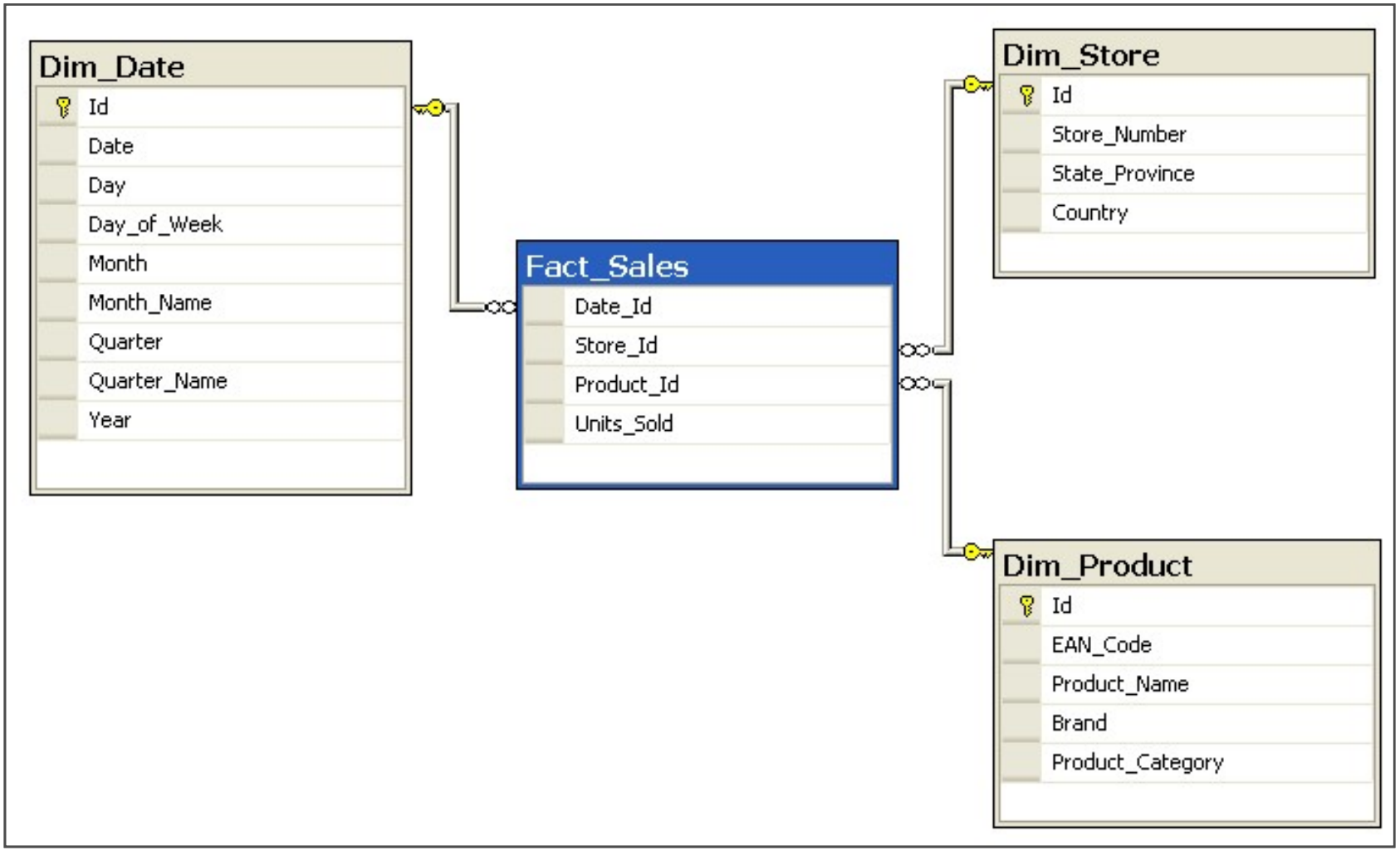
- 3NF or Third Normal Form
  which is frequently used in OLTP applications

Each model will be detailed at a high-level in the coming slides

Note: Each of the models listed is rigid in it's rules for data relationships and typically takes years to develop

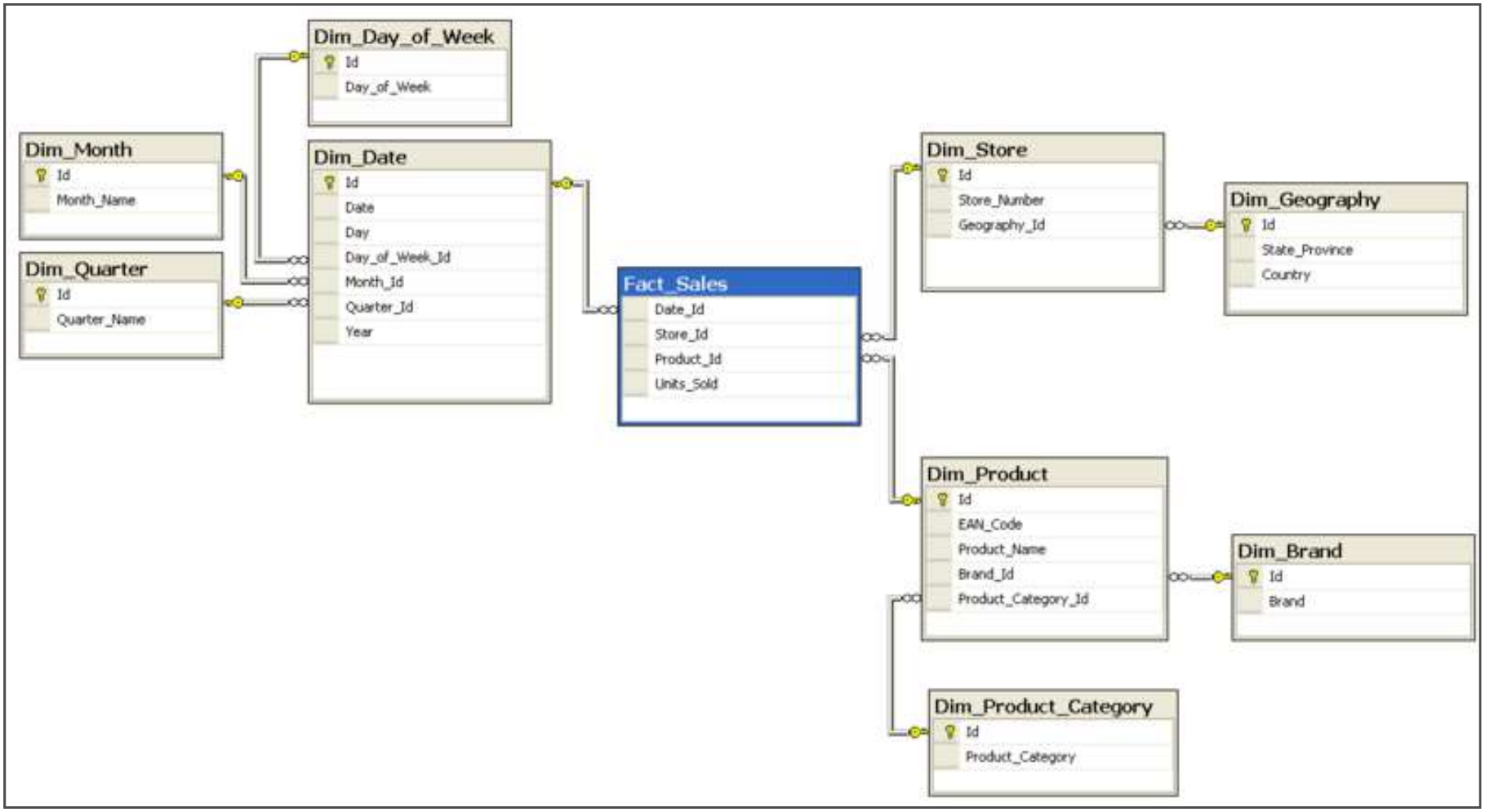
Pivotal

Star Schema Model

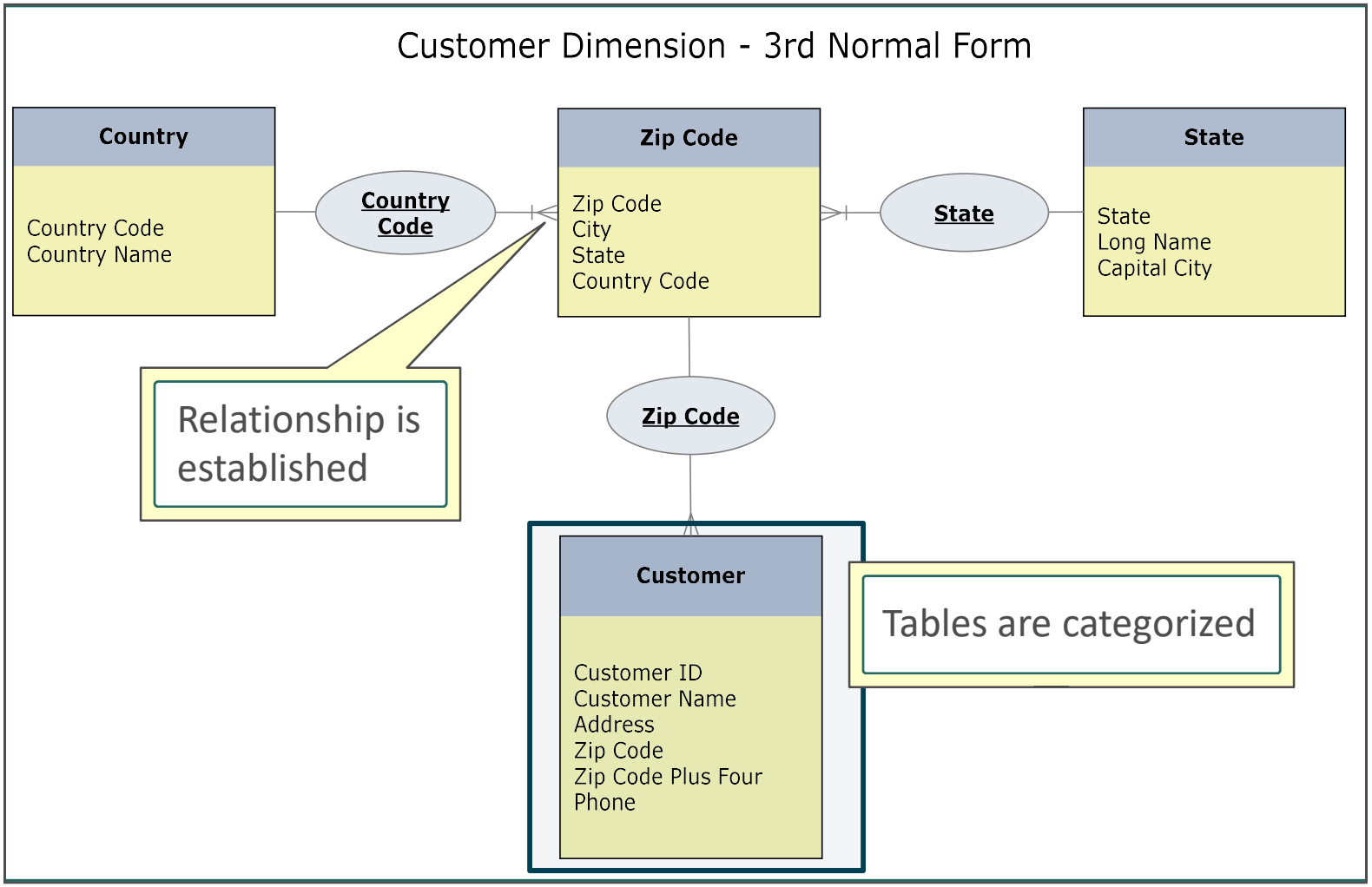


Pivotal

Snowflake Schema Model



3NF Model



A Pivotal Approach

Consider the following when modeling for Greenplum:

- Primary keys are essential for OLTP and as such their "primary-ness" should be established prior to loading
- Business users look at the data based on how they use it – Model to reflect this "use" pattern
- Don't blindly follow a pattern espoused by experts unless it fits the usage of the data
- Data reorganization within a greenplum system is exceedingly fast, so iterate over models until one resonates with the users
- Storage on MPP is not usually as constrained as on other systems so it may be appropriate and useful to have multiple copies of a data set to support different uses

Pivotal

A Pivotal Approach (cont'd.)

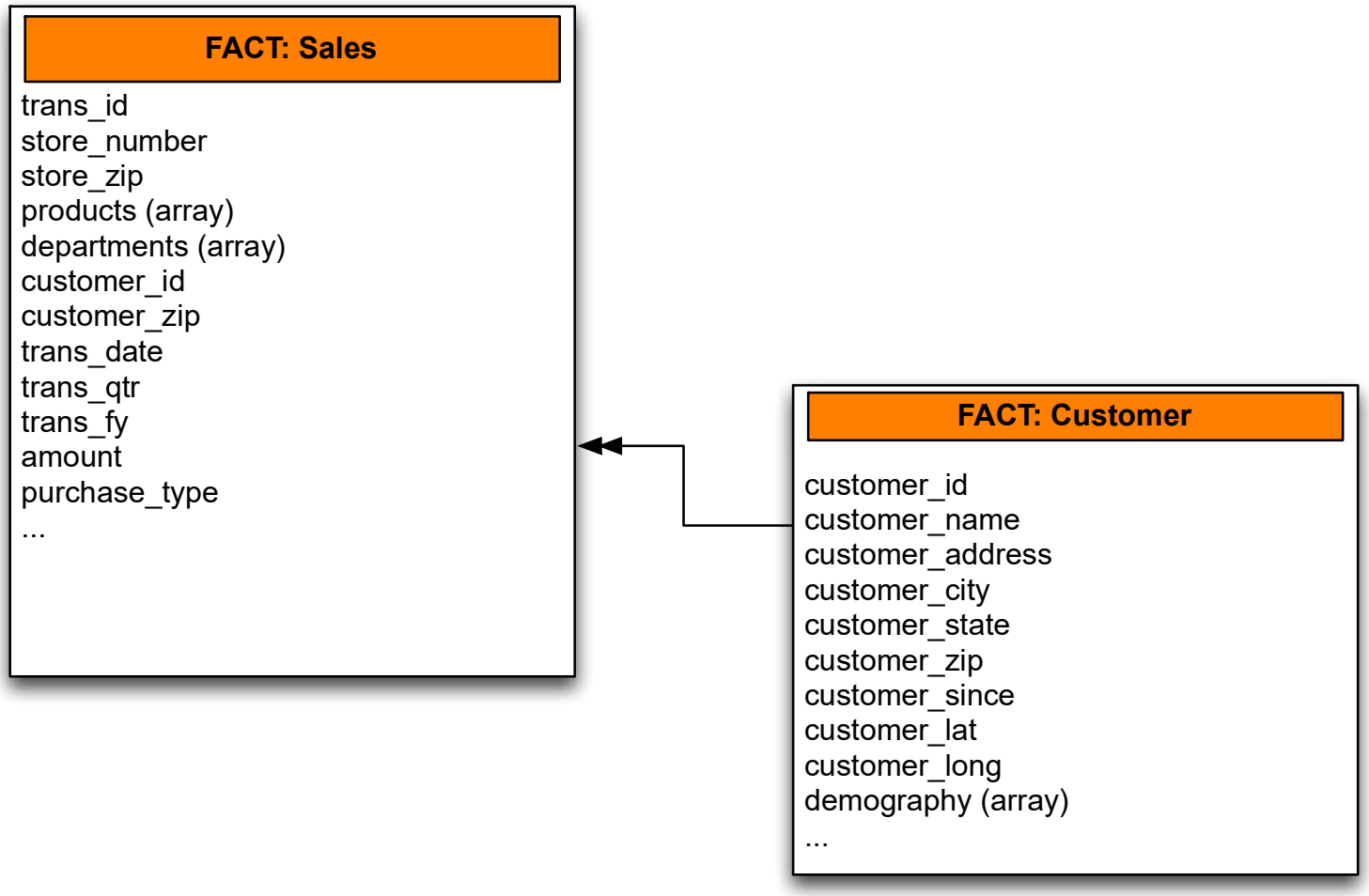
Other considerations when modeling for Greenplum:

- Avoid over-modeling which can result in large numbers of joins that tend to not be as performant
- Flattening is okay and encouraged in Greenplum and it has the side effect of making more sense to business users…
- Model, test, iterate – be agile and figure out what works best for the determined use
- It is possible and can be very functional to leave data outside the database, accessing it as needed via the powerful external table mechanism in Greenplum*

* Data can be filesystem based or can live in hadoop

Pivotal

Pivotal Approach Example



Pivotal Approach Example Explained

As suggested it can be very effective to create a flattened view of data:

- More closely fits with a business view of the data
- Easier to query
- Faster to query
- Can still join to information not flattened into the fact
  - e.g.
    Customer demographics
    Customer geo information
- Minimizes in-memory data structures

Pivotal Approach Possible Optimizations

- Don’t over-think it
- Use the power of the system to evaluate and test different approaches
- Don’t force data sets into a certain model
- Relax -- it is just data and can be reorganized into most any “shape” that is desired
- Utilize the right tool for the right job
  e.g. don’t use GPDB to enforce RI; don’t use your OLTP database for analytics
- Don’t forget the powerful capabilities of GPDB to ingest and utilize data that may be outside of the database
- Greenplum includes many features to improve performance: compression, columnar, partitions, etc.

Pivotal

Review

- Models & Background
- Pivotal Approach
- Possible Optimizations


Pivotal
A NEW PLATFORM FOR A NEW ERA