Design and implementation of open-data data warehouse Documentation

Release 0.0.1

Richard Banyi

May 09, 2016

Contents

Abstract

SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA FACULTY OF ELECTRICAL ENGINEERING AND INFORMATION TECHNOLOGY

Study programme Robotics and Cybernetics

Study field number 2647

Study field 9.2.7. Cybernetics

Training workplace Institute of Robotics and Cybernetics :Thesis supervisor: Ing. Ján Cigánek, PhD.

Thesis consultant Ing. Štefan Urbánek

CHAPTER 1

Introduction

1.1 Data

1.1.1 What is Data?

Data is all around us. But what exactly is? Data is a value assigned to a thing. What can we say about the balls in the picture? They are tennis balls, right? So one of the first data points we have is that they are used for tennis. Tennis is a category sport, so this helps us to to put the balls in a taxonomy. But there is more to them. We have the colour: "green", the condition "new". They all have sizes. All kind of objects have lot of data attached to them. Even people do: they have a name a date of birth, weight, height etc. All these things are data [1].

Fig. 1.1: Tennis Balls

Qualitative data is everything that refers to the quality of something. A description of experiences, a description of colours, and interview are all qualitative data. Data that can be observed but not measured.

Quantitative data is data that refers to a number. Data that can be measured. E.g. the number of tennis balls, the size, a score on a test etc.

Categorical data puts the item you are describing into a category. For example the condition "new", "used", "broken" etc.

Discrete data is a numerical data that has gaps in it: e.g. the count of golf balls. There can only be whole numbers of tennis balls, there is no such thins as 0.5 golf balls.

Continues data is a numerical data with a continues range: e.g. size of a tennis balls can be any size (86.02mm), or the size of your foot (as opposed to your shoe size, which is discrete). In continues data, all values are possible with no gaps between .

[1] Authors: School of Data Organization: School of Data Date: Sep 02, 2013 Available from: Data Fundamentals

1.1.2 From Data to information knowledge

Data, when collected and structured suddenly becomes a lot more useful.

Category	Contract
Date	2015
Amount	$1232.21
Recipient	Apple Inc.

Data: Content that is directly observable or verifiable; a fact - it's -5C outside.

Information: Content that represents analyzed data - "it's -5C outside I'll take a warm coat".

Knowledge: "I remember the last time when was this cold I got a cold. I'll therefore take a scarf and gloves. But first I'll check with Brian. He usually dress too light for this kind of weather." Knowledge is created when the information is learned, applied and understood.

Context: One thing incredibly important for data is context: A number or quality doesn't mean a thing if you don't give context. So explain what you are showing – explain how it is read, explain where the data comes from and explain what you did with it. If you give the proper context the conclusion should come right out of the data.

1.1.3 Finding Data

Data Source	Description
csv	Comma Separated values (CSV)
xls	MS Excel Spreadsheet
gdoc	Relational database table
mongodb	MongoDB database

There are three basic ways of getting hold of data:

1. **Finding data** - involves searching and finding data that has been already released
2. **Getting hold of more data** - asking for "new" data from official sources e.g. trough Freedom of Information requests. Sometimes data is public on a website but there is not a download link to get hold of it in bulk! This data can be liberated with what datawranglers call scraping.
3. **Collecting data yourself** - gathering data and entering it into a database or a spreadsheet.

1.1.4 Identify data source

In recent years *governments* have begun to release some of their data to the public - open data. Many governments host special (open) government data platforms for the data they create. For example the UK government started UK Data, or USA data.gov. Other sources of data are large *organisations*. The World Bank and the World Health Organization for example regularly release reports and data sets. Scientific projects and institutions release data to the scientific community and the general public. Open data is produced by NASA for example, and many specific disciplines have their own data repositories, some of which are open.

1.1.5 Problem with sources

There are plenty of places where you can get hand on open datasets which are open to public, however these sources are often unstructered, very messy, ambiguous, mis-use of class attributes, non-consistent, missing values, etc. The unstructured data growing quickiest than the other, and their exploitation could help in business decision.

```
5896 34513500 34513500-1     \N      \N Transport equipment
3031 48825000 48825000-7     \N      \N      Software package
5821 34144900 34144900-7     1.9721673495    \N      Transport equipment
```

1.2 Open Data

Do you know exactly how much of your tax money is spent on street lights or on public transportation? And what is in the air that you breathe along the way? Where in your region will you find the best job opportunities and the highest number of fruit trees per capita? When can you influence decisions about topics you deeply care about, and whom should you talk to?

New technologies now make it possible to build the services to answer these questions automatically. Much of the data you would need to answer these questions is generated by public bodies. However, often the data required is not yet available in a form which is easy to use - take for example our country Slovakia it still lack of data transparency and creating data sets which can be easy to used.

The notion of open data and specifically open government data - information, public or otherwise, which anyone is free to access and re-use for any purpose - has been around for some years.

1.2.1 Why open Data?

Open data, especially open government data, is a tremendous resource that is as yet largely untapped. Many individuals and organisations collect a broad range of different types of data in order to perform their tasks. Government is particularly significant in this respect, both because of the quantity and centrality of the data it collects, but also because most of that government data is public data by law, and therefore could be made open and made available for others to use. Why is that of interest?

There are also many different groups of people and organisations who can benefit from the availability of open data, including government itself. At the same time it is impossible to predict precisely how and where value will be created in the future.

It is already possible to point to a large number of areas where open government data is creating value. Some of these areas include:

- Transparency and democratic control
- Participation
- Self-empowerment
- Improved or new private products and services
- Innovation
- Improved efficiency of government services
- Improved effectiveness of government services
- Impact measurement of policies
- New knowledge from combined data sources and patterns in large data volumes

Open government data can also help you to make better decisions in your own life, or enable you to be more active in society. A woman in Denmark built findtoilet.dk, which showed all the Danish public toilets, so that people she knew with bladder problems can now trust themselves to go out more again. Services like ‘mapumental’ in the UK and ‘mapnificent’ in Germany allow you to find places to live, taking into account the duration of your commute to work, housing prices, and how beautiful an area is. All these examples use open government data.

Open data is also of value for government itself. For example, it can increase government efficiency. The Dutch Ministry of Education has published all of their education-related data online for re-use. Since then, the number of questions they receive has dropped, reducing work-load and costs, and the remaining questions are now also easier for civil servants to answer, because it is clear where the relevant data can be found. Open data is also making government more effective, which ultimately also reduces costs.

While there are numerous instances of the ways in which open data is already creating both social and economic value, we don't yet know what new things will become possible. New combinations of data can create new knowledge and insights, which can lead to whole new fields of application. We have seen this in the past, for example when Dr. Snow discovered the relationship between drinking water pollution and cholera in London in the 19th century, by combining data about cholera deaths with the location of water wells.

This untapped potential can be unleashed if we turn public government data into open data. This will only happen, however, if it is really open, i.e. if there are no restrictions (legal, financial or technological) to its re-use by others. Every restriction will exclude people from re-using the public data, and make it harder to find valuable ways of doing that. For the potential to be realized, public data needs to be open data.

1.2.2 What is Open Data?

Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike [3].

The full Open Definition gives precise details as to what this means. To summarize the most important:

- **Availability and Access:** the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- **Re-use and Redistribution:** the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
- **Universal Participation:** everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed.

If you're wondering why it is so important to be clear about what open means and why this definition is used, there's a simple answer: **interoperability**.

Interoperability denotes the ability of diverse systems and organizations to work together (inter-operate). In this case, it is the ability to interoperate - or intermix - different datasets.

Interoperability is important because it allows for different components to work together. This ability to componentize and to 'plug together' components is essential to building large, complex systems. Without interoperability this becomes near impossible — as evidenced in the most famous myth of the Tower of Babel where the (in)ability to communicate (to interoperate) resulted in the complete breakdown of the tower-building effort. We face a similar situation with regard to data. The core of a "commons" of data (or code) is that one piece of "open" material contained therein can be freely intermixed with other "open" material. This interoperability is absolutely key to realizing the main practical benefits of "openness": the dramatically enhanced ability to combine different datasets together and thereby to develop more and better products and services.

Providing a clear definition of openness ensures that when you get two open datasets from two different sources, you will be able to combine them together, and it ensures that we avoid our own 'tower of babel': lots of datasets but little or no ability to combine them together into the larger systems where the real value lies [4].

[3] Open Definition see
[4] Open Knowledge Open Data Handbook

1.3 Data Warehouse Fundamentals

1.3.1 What is Data Warehouse

A data warehouse (DW or DWH) is a system used for reporting and data analysis. Data Warehouse's are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analyses.

A data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making [5].

1.3.2 Operational Systems

An operational system directly supports the execution of a business process. By capturing details about significant events or transactions. A sales system, for example captures information about orders, shipments, and returns.

Operational systems must enable several types of database interaction, including inserts, updates, and deletes - these interactions are almost always atomic. For example, an order entry system must provide for the management of lists of products, customers, and salespeople; the entering of orders; the printing of order summaries, invoices, and packing lists; and the tracking order status. The operational system is likely to update as things change (if a customer moves, his/her old address is no longer useful so it is simply overwritten), and archive data ones it's operational usefulness has ended. Operational systems are implemented in a relational database, the design may called entity-relationship model, or ER model. The schema of operational systems are highly accepted to be in third normal form.

1.3.3 Analytic Systems

An analytical system supports the *evaluation* of a business process. How are orders trending this month versus last? Where does this put us in comparison to our sales goals for the quarter? Is a particular marketing promotion having an impact on sales? Who are our best customers?

Interaction with an analytic system takes place through queries that retrieve data about business processes. Historic data will remain important to the analytic system long after its operational use has passed.

OPERATIONAL SYSTEM VS. ANALYTICAL SYSTEM

	Operational System	Analytic System
Purpose	Execution of a business process	Measurement of a business process
Primary Interaction Style	Insert, Update, Delete, Query	Query
Scope of Interaction	Individual transaction	Aggregated transactions
Querry Patterns	Predictable and stable	Unpredictable and changing
Temporal Focus	Current	Current and historic
Design Optimaziation	Update concurrency	High-performance query
Design Principle	Entity-relationship (ER) design in third normal form (3NF)	Dimensional design (Starschema or Cube)
Also Known As	Transaction System,Online Transaction Processing System (OLTP),Source System	Data Warehouse System,Data Mart

[5] The Data Warehouse ETL Toolkit, Ralph Kimball, Joe Casetra, Copyright 2004 by Wiley Publishing, Inc. All rights reserved., eISBN: 0-764-57923-1

1.3.4 Analytic Databases and Dimensional Design

The dimensional model of a business process is made up of two components: *measurements* and their *context*. Known as facts and dimensions, these components are organized into a database design that facilities a wide variety of analytic usage. Implemented in a relational database, the dimensional model is called a star schema. Implemented in a multidimensional database, it is known as a cube. The core of every dimensional model is a set of business metrics that captures how a process is evaluated, and a description of the context of every measurement [6].

Purpose

Analytic systems and operational systems serve fundamentally different purposes. An operational system supports the execution of a business process, while and analytic system supports the evaluation of the process [7].

Measurement and Context

Dimensional design supports analysis of a business process by modeling how it is measured. Consider the following business questions:

- What are gross margins by product category for June?
- What is the average transaction by states level?
- What is the return rate by visitors?

These questions do not focus on individual activities or transactions. To answer them, it is necessary to look at a group of transactions - in a bigger picture. Each of these questions reveals something about how its respective business process is measured.

Every dimensional solution describes a process by capturing what is measured and the context in which the measurements are evaluated [8].

Facts and Dimensions

In a dimensional design, measurements are called facts, and context descriptors are called dimensions. Facts tend to be numeric in value. Elements that are aggregated, summarized, or subtotaled are facts.

FACTS	DIMENSIONS
Amount	Product
Min Amount	Agency
Max Amount	Award
	Geography

1.3.5 The Star Schema

A dimensional design for a relational database is called a star schema. Related dimensions are grouped as columns in dimension tables, and the facts are stored as columns in a fact table.

Dimension tables are not in third normal form. A dimensional model serves a different purpose from ER model. It is not necessary to isolate repeating values in an environment that doesn't support transaction processing. When additional normalization is performed within dimensions, in such cases, the schema is referred as a snowflake.

Dimension Tables

In a star schema, a dimension table contains columns representing dimensions. These columns provide context for facts.

[6] Excerpt From: Adamson, Christopher. "Star Schema The Complete Reference." Copyright 2010 by The McGraw-Hill Companies, Inc. All rights reserved. ISBN: 978-0-07-174433-1

[7] Excerpt From: Adamson, Christopher. "Star Schema The Complete Reference." Copyright 2010 by The McGraw-Hill Companies, Inc. All rights reserved. ISBN: 978-0-07-174433-1

[8] Excerpt From: Adamson, Christopher. "Star Schema The Complete Reference." Copyright 2010 by The McGraw-Hill Companies, Inc. All rights reserved. ISBN: 978-0-07-174433-1

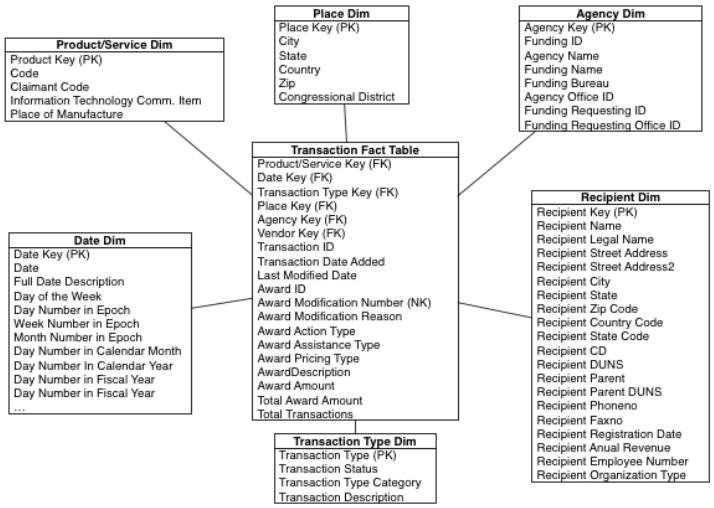

Fig. 1.2: Star Schema example

Fact Table

At the core of a star schema is the fact table. Each row in the fact table stores facts at a specific level of detail. This level of detail is known as the fact table's grain

CHAPTER 2

Building The Data Warehouse

2.1 Plan

2.1.1 Goals of Data Warehouse

Before we delve into the details of dimensional modeling and implementation, it is helpful to focus on fundamental goals of the data warehouse. How can we focus on these these fundamental goals if we are missing the most important thing, the **data**.

Based on our experience, the data is the universum that drive the bedrock requirements for the data warehouse. The data comes first, than the technology and the bussiness model.

Finding the Data

After couple of weeks of searching we have identified several open to public data sources:

- UK Gov
- Open Spending
- U.S. Government's open data
- Usa Spending Gov

We have decided to looking for government data. We exemined couple of datasets and we picked Usa Spending Gov. Their data looked the most promicing and they provide great document information about the data fields, description and their formats.

Identifying the data source

All the prime recipient transaction data on *USAspending.gov <https://www.usaspending.gov>* is reported by the federal agencies making contract, grant, loan, and other financial assistance awards. After identifying the data source and examination of the data sets we have decided to use this data for our business model and data warehouse model.

Mission of data warehouse

We have concluded the following goals for our data warehouse:

The main mission of the data warehouse is to publish the federal organisations data. The key success of our data warehouse is whether the data warehouse effectively contributes to the general public. The success of a data warehouse begins end ends with its users.

The data warehouse is going to be open to public. **It must make the information easily accessible.** The content of the data warehouse must be understandable. The data must be intuitive and obvious to the business user, not merely the developer.

The data warehouse must present the federal organisations information consistently. Data must be carefully assembled from a variety of sources around the organization, cleansed, quality assured, and released only when it is fit for user consumption.

The data warehouse must be adaptive and resilient to change. We want to track changes of federal awards made by federal agencies. The data warehouse must be designed to handle this change.

2.2 Data warehouse Environment

It is helpful to understand the pieces of the data warehouse before we begin to combine them. Each component serves specific function.

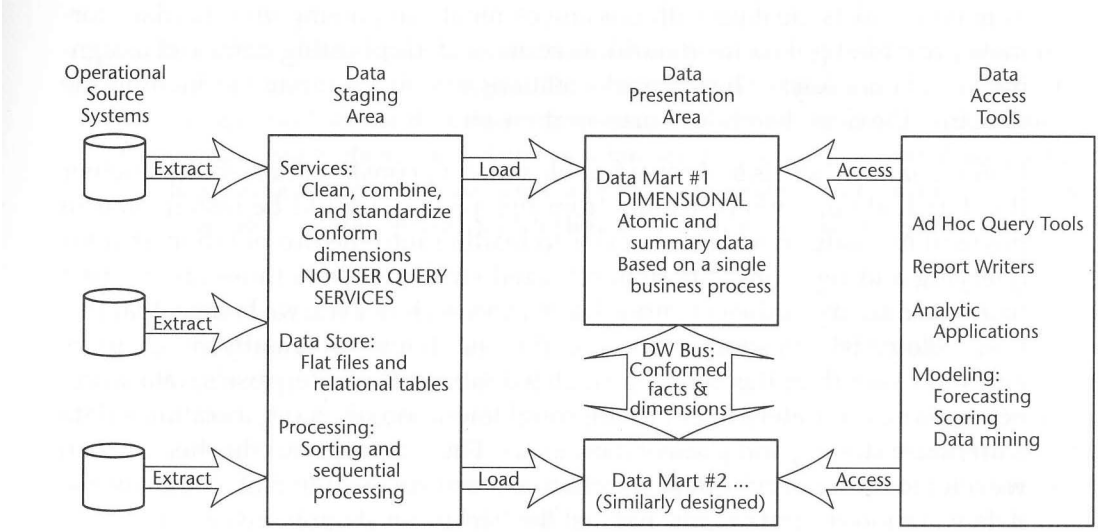

Figure 1.1 Basic elements of the data warehouse.

Fig. 2.1: Data warehouse components

2.2.1 Operational Source systems

The operational system directly captures the execution of a business process, in our case capture the transactions of federal agencies. The operational source system is outside the data warehouse, we have no control over the content and format of the data in these systems. In our case the operational system is usaspending.gov where we just download directly from they site a .csv flat file.

2.2.2 Data Staging Area

The data staging is a physical storage area and a set of extract-transformation-load (ETL) jobs. The data staging are is everything between the source systems and the data presentation area. This is the stage where we perform various process on the data to fit into data warehouse environment.

2.2.3 Data presentation

The data presentation area is where the structured data is organized, stored, and made available for querying and for analytical applications. The presentation area is based on online analytic processing (OLAP) technology, this means the data is stored in cubes.

2.2.4 Data access tools

The final piece of the data warehouse is the data access tool. Obviously the first access tool can be a simple query tool for querying. We provide to our end users the general public set of tools for development of reporting, analysis and browsing of data trough our web application using the concept of Cubes.

2.3 Data Staging Area

The ETL system is the foundation of the data warehouse. We have design an ETL system that extracts data from the source system, enforces data quality and consistency, conforms data so that separate flat files can be used together, and delivers data into presentation layer where we build the application so that end users can make decisions. We haven't used any ETL tools every process/job is all hand coded.

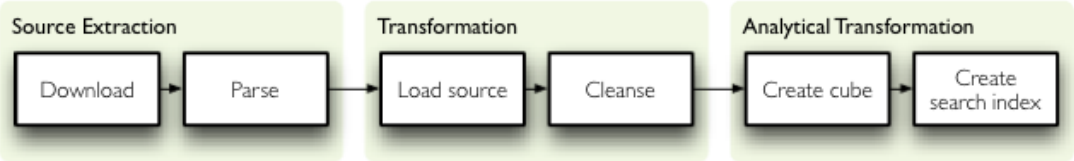

Fig. 2.2: ETL process flow

2.3.1 Extract - E

The usaspending.gov website doesn't provide any API, only single page with the download able link. The raw data coming from the source system is stored locally on the disk. We have downloaded 4 different .csv flat files.

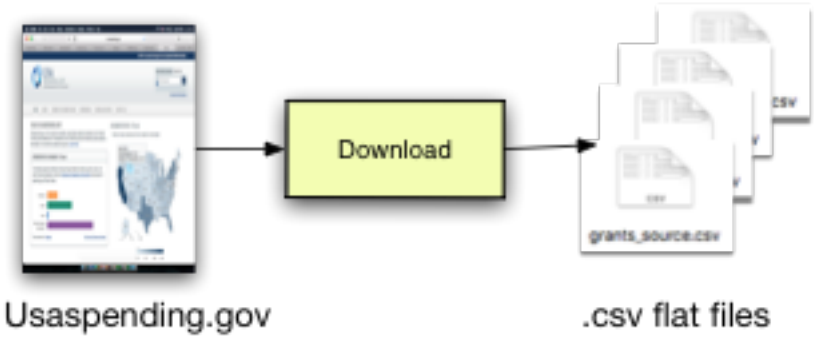

Fig. 2.3: Figure [5]

2.4 Dimensional Modeling

2.4.1 Business Process

The first step in the design is to decide what business process to model by understanding of the business requirements with an understanding of the available data. In our open government case study, the general public wants to better understand how the government spends money, what kind of transactions are made in their neighbourhood. Thus our process is a transactional model. This transactional data will allow us to analyse what kind of awards are made by federal agencies in which states on what days and to whom. Brief description of business model that we'll use in our case study to make dimension and fact tables more understandable. Imagine a federal agency for example Department of Agriculture making a award for a recipient 1901 Combine Group, LLC for a combine harvester in Texas on 2015. To summarise it WHICH federal agency is awarding WHOM for WHAT and WHERE is the place of the performance of the transaction made.

2.4.2 Declare the Grain

Once the business process has been identified. we faced a serious decision about the granularity of the data warehouse. What level of data detail should be made available in the dimensional model? After identifying the data, we had couple of options to choose. We wanted tackling the data at it's lowest level, most atomic grain made the most sense. The more detailed and atomic the fact measurement, the more things we know for sure about federal awards. In this regard, atomic data was the perfect match for the dimensional approach. Atomic data provides the maximum analytic flexibility because it can be constrained and rolled up in every way possible. In our case study, the most granular data is an individual transaction made by federal agencies. Because of this level of grain we ensured maximum dimensionality and flexibility. Providing access to the transactions information gave us very detailed look at federal award changes. For example, the end users want to see how many transaction were made for one individual award or how the award has changed over period of time, if the agency made a modification to an award, reduced a portion of the original award amount or made additional funding.None of them could have been answered if we wouldn't elected the lowest granularity just the summarised data.

2.4.3 Choose the Dimensions

After we have declared the grain of the fact table, the recipient, agency, date, geography, award, dimensions fall out immediately. We assume that the calendar date is the date when the award was signed.

In our case study we have decided on the following dimensions:

Dimension tables are not in third normal form. A dimensional model serves a different purpose from ER model. It wasn't necessary to isolate repeating values in an environment that doesn't support transaction processing. If we would have made additional normalisation within dimensions, we would end up with the schema that is referred as a snowflake. We have encouraged to resist the urge to snowflake given our to primary design, ease of use and performance.

- Snowflaked tables makes for much more complex presentation.
- Database design will struggle with the complexity of the snowflaked schema.
- Numerous tables and joins usually translate into slower query performance.
- Minor disk space savings.
- Snowflaking slows down the user's ability to browse within the dimension.

Dimension tables also contain key columns that uniquely identify something in an operational system. These key columns are referred to as natural keys. The separation of surrogate keys and natural keys allows the data warehouse to track changes, even if the originating operational system does not.

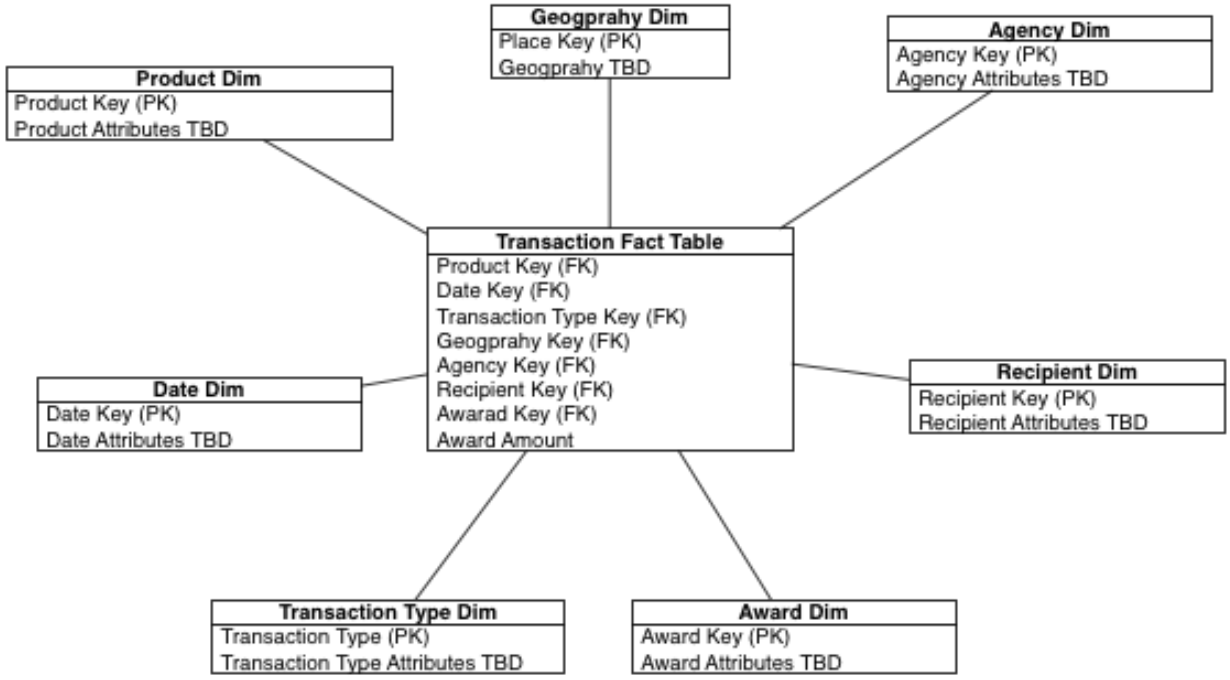

Fig. 2.4: Preliminary star schema.

2.4.4 Identify Facts

At the core of a star schema is the fact table. In addition to presenting the facts, the fact table includes surrogate keys that refer to each of the associated dimension tables. Each row in the fact table stores facts at a specific level of detail of our grain that we have declared. Facts tend to be numeric in value. We have made the decision that the award amount is going to be our fact measurement which will appear in our fact table. We have decided to stored physically in the data warehouse only one fact the award mount, which is additive across all dimensions.

2.4.5 Suroggate Keys

In the star schema, each dimension table is given a surrogate key. This column is a unique identifier, created exclusively for the data warehouse. The surrogate key is the primary key of the dimension table. In our case, surrogate keys are randomly generated integers that are assigned sequentially when populate dimension tables during the ETL process. For example, the first recipient record is assigned a recipient surrogate key with the value of 1, the next recipient record is assigned recipient key with value 2, and so forth. The surrogate keys serve to join the dimension tables to the fact table. One of the most important reasons why are we using surrogate keys and doesn't just rely on natural keys from the source system is to support handling changes to dimension table attributes.

2.5 Dimensional Table Attributes

CHAPTER 3

References

[1] Authors: School of Data, Organization: School of Data Date: Sep 02, 2013 Available from: Data Fundamentals