The Universal Data Cube Model

Curran Kelleher

The Universal Data Cube theoretical model (UDC model) is a conceptual framework which characterizes distributed hierarchical data cubes in terms of their content, structure, and metadata. While this model is theoretical, it is designed specifically to solve many concrete and ubiquitous problems in data publishing, navigation, interactive visualization and analysis:

- No open standard for publishing and consuming hierarchical data cubes with rich metadata exists.
- The metadata necessary to determine whether two data sources provide comparable data is often not machine readable, therefore merging of data sources requires extensive manual effort.
- Data sources using different coding schemes for identifying records require manual work (creation of code translation tables) to merge and compare.
- Data sources representing numeric values with the same meaning using different units require manual work (creation of unit conversion functions) to merge and compare.
- The data infrastructure for many visualization tools must still often be developed from scratch, and is domain or task specific.
- Hierarchical and tabular data are widely considered findamentally different kinds of data, and tools are often developed for only one or the other.
- The work required to create a tool in which users can navigate categorical or regional hierarchies is repeated for every such tool created.

All these problems are solved effectively by the UDC model.

A data cube is an a form of data abstraction which is comprised of *dimensions* (containing objects) and *measures* (representing properties). A *hierarchical data cube* is a data cube whose dimensions may be hierarchical. The UDC model is an extention of the hierarchical data cube concept which aims to provide a comprehensive framework for describing and using data cubes. We first introduce the structural terms used in the UDC Model, partitioned into those relating to dimensions and those relating to measures. We then then discuss how these terms can be composed to form the conceptual framework for a distributed data system potentially containing all possible hierarchical data cubes, namely the *universal data cube*. We conclude with a discussion of how this structure can be practically used for data publishing, navigation, interactive visualization and analysis.

1 Dimensions

A *record* represents a region of time, a region of space, an object category or an individual object. Example records include the year "1990" (a region of time), the country "USA" (a region of space), the industry sector "Mining" (an object category), and the iris flower with id "25" (an individual object). Records can be hierarchical. Each record may have a *parent record*. For example, the parent record of the US state "Massachusetts" may be the country "USA". For a given record, the records which have it as a parent are considered its *child records*.

Each record can have more than one parent. Therefore the structure of record relationships is not limited to a hierarchy (tree), but in general is a *topology* (graph). A record a is considered a parent of another record b if and only if the region, category or individual represented by b is fully contained within that represented by a. Many *record hierarchies* (containment trees) can be expressed within a single *record topology* (containment graph).

A *record product* can be defined for regional or categorical records, which represents a composite region. For example, the product of the country "USA" and the year "1990" defines simultaneously a region of time and space, namely "The USA in the year 1990".

A *level* represents a class of records which are on the same level in a record hierarchy. Example levels include "Year", "Country", "Industry Sector" and "Iris". Levels can be hierarchical. Each level may have a *parent*

level. For example, the parent record of the level "US state" may be the level "Country". For a given level, the levels which have it as a parent are considered its *child levels*. A level tree represents a record topology. A path from the root to a leaf of a level tree (or any subpath thereof) represents a record hierarchy.

A *dimension* represents a set of levels which could all potentially belong to the same level tree, and the record topology represented by these levels. Example dimensions include "Time", "Space", "Industry" and "Iris Category". All records are contained within levels, and all levels are contained within dimensions.

2 Measures

A *quantity* is a kind of numeric property. Example quantities include Currency, Quantity of People, Mass, and Speed. A *unit* is a concrete realization of a quantity. Example units include US Dollars, Thousands of People, Kilograms, and Kilometers per Hour. An *aggregation operator* defines a method of aggregating numeric values. Example aggregation operators include "Sum" and "Average". A *measure* is a numeric property of records or products thereof, defined by a quantity and an aggregation operator. Here are some example measures, shown in the form: "measure name", "quantity", "aggregation operator":

- "Average Income", "Currency", "Average"
- "Population", "Quantity of People", "Sum"
- "Employment", "Quantity of People", "Sum"
- "Average Speed Limit", "Speed", "Average"

3 Data

A *data cube* in general is a mapping from record products to measure values. For example, a data cube of the US Census data would contain a mapping from the record product "The USA in the year 1990" to a value for the measure "Population" in a specific unit.

A *dimension instance* is a specific subset of records from one level of a dimension. A *measure instance* is a concrete realization of a measure annotated with its unit, in other words a $(measure, unit)$ pair. The structure of a data cube is defined by a set of dimension instances and a set of measure instances. The *cells* of a data cube are defined by all possible record products in which one record is taken from each of the data cube's dimension instances. The content of data cube is defined by the mapping from its cells to numeric values for each of its measure instances.

A *dataset* is a collection of data cubes. A dataset whose data cubes all contain the same set of measures instances and contain levels representative of the same set of dimensions can be interpreted as a data cube in which each dimension is hierarchical. Such a data set is called a *hierarchical data cube*.

4 Properties

The UDC model can be cleanly divided into realms of *knowledge* and *data*. Knowledge in this context refers to the terminology potentially used across data sets: records, levels, dimensions, quantities, units, and measures. Data in this context refers to actual data set contents and their descriptors: dimension instances, measure instances, data cubes and data sets.

5 Auxiliary Data

Code tables are mappings from strings of a specific coding scheme to record URIs. *Unit conversion tables* are mappings from unit pairs to conversion factors.

6 Operations

The knowledge realm of the UDC model can be represented as a semantic graph (a graph with labeled edges). Any semantic graph query on this knowledge base is possible.

Data cubes can be projected (queried) by selecting a subset of their records and measures. We use the notion of a *record selection* for describing subsets of records used when specifying a projection. A record selection in a

given dimension instance can be defined by a single record representing the root of a record subtree. The records at the level represented by the given dimension instance are selected.

Data cubes are comparable when their set of dimension instances represent the same set of levels, and their set of measure instances represent the same set of measures. Comparable data cubes can be merged by taking the union of records in their respective dimension instances, and harmonizing (converting) units of comparable measures when necessary (using unit conversion tables).