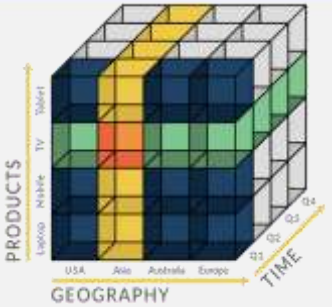
CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud: Datacubes

Prof. Reza Farivar

Datacube Origins

- The OLAP cube grew out of a simple idea in programming: take multi-dimensional data and put it into what is known as a ‘2-dimensional array’ — that is, a list of lists
- A Datacube is a data structure
  - A sophisticated nested array
  - Compressions schemes
  - Data aggregation techniques when the cube outstrips the host's memory
- What if you have a massive dataset and want to run queries
  - Real technological constraints lead to the creation of the OLAP cube
  - Cache subsets of data within the nested array — and occasionally persist parts of the nested array to disk
- Today, ‘OLAP cubes’ refer specifically to contexts in which these data structures far outstrip the size of the hosting computer’s main memory — examples include multi-terabyte datasets and time-series of image data



Datacube Schemas

- OLAP cube requires that data teams manage complicated pipelines to transform data from an SQL database into cubes
- If you were working with a large amount of data, such transformation tasks could take a long time to complete, so a common practice would be to run all ETL (extract-transform-load) pipelines before the analysts came in to work.
- Using OLAP cubes in this manner also meant that SQL databases and data warehouses had to be organized in away that made for easier cube creation

Datacube Dimensional Modeling

- Early practitioners observed that certain access patterns occurred in every business
  - Kimball, Inmon and their peers
- They developed repeatable methods to turn business reporting requirements into data warehouse designs
  - designs that allow teams to extract the data they need in the formats they need for their OLAP cubes
- If you became a data analyst in the previous two decade, you had to “model” your data according to these best practices
  - Kimball dimensional modeling, Inmon-style entity-relationship modeling, or data vault modeling
  - Methods for organizing the data in the data warehouse to match the businesses' analytical requirements

Datacube Operations

- Slicing: the act of picking a rectangular subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension
- Dicing: produces a subcube by allowing the analyst to pick specific values of multiple dimensions
- Drill Up / Down: allows the user to navigate among levels of data ranging from the most summarized (up) to the most detailed (down)
- Roll-up: A roll-up involves summarizing the data along a dimension
- Pivot: allows an analyst to rotate the cube in space to see its various faces