CLOUD COMPUTING APPLICATIONS

Analytics in the Cloud: Rise and Fall of Datacubes

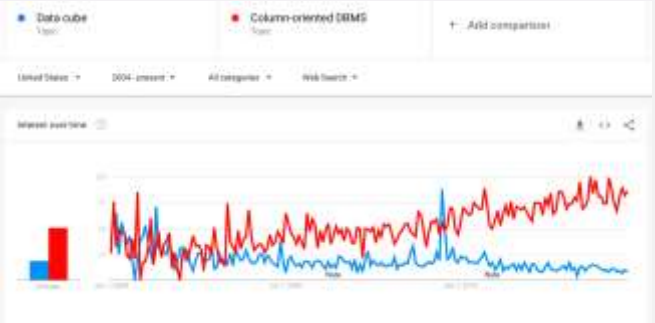
Prof. Reza Farivar

Datacube vs. Columnar RDBMS

- OLAP cubes traditionally known for extreme performance advantage over row-oriented RDBMS
  - Less important with recent advances in computers and columnar storage
- OLAP cubes demand that you load a subset of the dimensions you're interested in into the cube
- Columnar databases allow performing similar OLAP-type workloads at equally good performance levels without the requirement to extract and build new cubes
- Note: OLAP Datacubes typically offer richer analysis capabilities than RDBMSs, which are limited by the constraints of SQL
  - The main justification Datacubes are still relevant

Current state

- Smaller companies are less likely to consider data-cube-oriented tools or workloads, and strict dimensional modeling has become less important over time
- Large tech giants (Google, Facebook, Amazon) have chosen columnar stores
  - Big Query, Redshift
- → One of the biggest shifts in data analytics over the past decade (2010 to 2020) is the move *away* from building Datacubes, to running OLAP workloads directly on columnar databases



Datacubes in the Future

- OLAP Datacubes typically offer richer analysis capabilities than RDBMSs, which are limited by the constraints of SQL
  - The main justification Datacubes are still relevant
  - OLAP cubes are being pushed upmarket
  - *We may return to them in the future*
- Example: Apache Kylin
  - Contributed by eBay in 2015
  - Build Datacubes on Hadoop and Spark
  - Utilizing HBase as Storage
  - Query billions of rows at sub-second latency
    - Identify a Star/Snowflake Schema on Hadoop
    - Build Cube from the identified tables
    - Query using ANSI-SQL and get results in sub-second, via ODBC, JDBC or RESTful API
- Druid, Apache Pinot (from Linkedin)
- Uber building a solution on Pinot + Presto