

**Datum: Managing Data Purchasing and Data Placement in a Geo-Distributed Data Market**

**Abstract:**

This paper studies two design tasks faced by a geodistributed cloud data market: which data to purchase (data purchasing) and where to place/replicate the data for delivery (data placement). We show that the joint problem of data purchasing and data placement within a cloud data market can be viewed as a facility location problem and is thus NP-hard. However, we give a provably optimal algorithm for the case of a data market made up of a single data center and then generalize the structure from the single data center setting in order to develop a near-optimal, polynomial-time algorithm for a geo-distributed data market. The resulting design, Datum, decomposes the joint purchasing and placement problem into two subproblems, one for data purchasing and one for data placement, using a transformation of the underlying bandwidth costs. We show, via a case study, that Datum is near optimal (within 1.6%) in practical settings.

**Existing System:**

Current data markets provide a variety of basic services: (i) *aggregation* of data from multiple sources, (ii) *cleaning* of data to ensure quality across sources, (iii) *ease of use*, through a unified API, and (iv) *low-latency delivery* through a geographically distributed content distribution network. As these market places mature they are increasingly adding other services as well. Besides providing raw data to clients, it is an inevitable trend for data markets to carry out value-added services built upon the data, such as analytics and machine learning APIs. Given the recent emergence of data markets, there are widely differing designs in the marketplace today, especially with respect to pricing.

Sets prices with a subscription model that allows a maximum number of queries (API calls) per month and limits the size of records that can be returned for a single query. Other data markets, e.g., Google BigQuery and Infochimps allow payments per query or per data set. In nearly all cases, the data provider and the data market operator each then get a share of the fees paid by the clients, though

how this share is arrived at can differ dramatically across data markets. The task of pricing is made even more challenging when one considers that clients may be interested in data with differing levels of precision/quality and privacy may be a concern.

**Proposed System:**

We initiate the study of jointly optimizing data purchasing and data placement decisions in geo-distributed data markets.

We prove that the task of jointly optimizing data purchasing and data placement decisions is NP-hard and can be equivalently viewed as a facility location problem.

We provide an exact algorithm with polynomial running time for the case of a data market with a single data center.