Joint Data Purchasing and Data Placement in a Geo-Distributed Data Market

Xiaoqi Ren, Palma London, Juba Ziani, Adam Wierman*
California Institute of Technology
{xren, plondon, jziani, adamw}@caltech.edu

ABSTRACT

This paper studies design challenges faced by a geo-distributed cloud data market: which data to purchase (data purchasing) and where to place/replicate the data (data placement). We show that the joint problem of data purchasing and data placement within a cloud data market is NP-hard in general. 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 and propose Datum, a near-optimal, polynomial-time algorithm for a geo-distributed data market.

1. INTRODUCTION

Ten years ago computing infrastructure was a *commodity*. Now, computing power and memory are *services* that can be cheaply subscribed to and scaled as needed via cloud providers like Amazon EC2, Microsoft Azure, etc.

We are beginning the same transition with respect to *data*. Data is broadly being gathered, bought, and sold in various marketplaces. However, it is still a commodity, often obtained through offline negotiations between providers and companies. Acquiring data is now one of the key bottlenecks for new tech startups.

This is beginning to change with the emergence of *cloud data markets*, which offer a single, logically centralized point for buying and selling data. Multiple data markets have recently emerged in the cloud, e.g., Microsoft Azure DataMarket [1], Factual [2], InfoChimps [3], Xignite [4], IUPHAR [5], etc. A rich literature has studied on cloud data market pricing. **This paper focuses on the engineering side of the design of a data market**, which has been ignored to this point. Supposing that prices are given, there are two important challenges that remain for the operation of a data market: 1) *data purchasing:* given prices and contracts offered by data providers, which providers should a data market purchase from to satisfy a set of client queries with minimal cost? 2) *data placement:* How should purchased data be stored and replicated throughout a geo-distributed data market in order to minimize bandwidth and latency costs? And which clients should be served from which replicas given the locations and data requirements of the clients?

In this paper, we present Datum, which jointly optimizes data purchasing and data placement costs for a geo-distributed data market. Datum first optimizes data purchasing as if the data market was made up of a single data center (given carefully designed "transformed" costs) and then, given the data purchasing decisions, optimizes data placement/replication. The "transformed" costs are designed to allow an architectural decomposition of the joint problem into subproblems that manage data purchasing (external operations of the data market) and data placement (internal operations of the data market). This decomposition is of crucial operational importance because it means that internal placement and routing decisions can proceed without factoring in data purchasing costs, mimicking operational structures of geo-distributed analytics systems today.

We have evaluated Datum using a case study, which shows that Datum is near-optimal (within 1.6%) in practical settings. Further, the performance of Datum improves upon approaches that neglect data purchasing decisions by $> 45\%$. Details are reported in the full version of this paper [6].

2. MODELING THE DATA CLOUD

We consider a setting where there are P data providers selling different data to C clients. Each data provider offers a set of quality levels $\mathcal{L}$. We use $q(l,p)$ to denote the data quality level l, offered by data provider p and use $f(l,p)$ to denote the fee charged by provider p for data of quality level l. Each client c sends a *query* to the data center, requesting particular data from multiple data providers. Denote the set of data providers required by the request from client query c by $G(c)$. The client query also specifies a minimum desired quality level, $w_c(p)$, for each data provider p it requests.

The data cloud in this marketplace is an aggregator and intermediary. We model the data cloud as a geographically distributed cloud consisting of D data centers. Each data center aggregates data from geographically separate local data providers, and data from data providers may be (and often is) replicated across multiple data centers within the data cloud. Denote the cost to transfer data of quality $q(l,p)$, originating from data provider p, from data center d to client c by $\alpha_{d,c}(l,p)$. And denote the cost to transfer data of quality $q(l,p)$ from data provider p to data center d by $\beta_{p,d}(l)$. Define binary variable $x_{d,c}(l,p)$ such that

*This work is partially supported by NSF grants CNS-1254169, CNS-1319820, NETS-1518941, and BSF grant 2012348.

SIGMETRICS '16 June 14-18, 2016, Antibes Juan-Les-Pins, France

ACM ISBN 978-1-4503-4266-7/16/06.
DOI: http://dx.doi.org/10.1145/2896377.2901486

$x_{d,c}(l,p) = 1$ if and only if data of quality $q(l,p)$, originating from data provider p, is transferred from data center d to client c. Define binary variable $y_{p,d}(l)$ such that $y_{p,d}(l) = 1$ if and only if data of quality $q(l,p)$ is transferred from data provider p to data center d.

Our goal is to provide a design that minimizes the operational costs of a data cloud. These costs include:

1) The *operation cost* due to transferring data of all quality levels from data providers to data centers.

$$\text{OperCost} = \sum_{p=1}^{P}\sum_{l=1}^{L_p}\sum_{d=1}^{D}\beta_{p,d}(l)y_{p,d}(l).$$

2) The *execution cost* due to transferring data of all quality levels from data centers to clients.

$$\text{ExecCost} = \sum_{c=1}^{C}\sum_{p\in G(c)}\sum_{l=1}^{L_p}\sum_{d=1}^{D}\alpha_{d,c}(l,p)x_{d,c}(l,p).$$

3) The *purchasing cost* (PurchCost) due to buying data from the data provider. Due to space limit, we only discuss the widely adopted per-query data contracting model here.

$$\text{PurchCost} = \sum_{c=1}^{C}\sum_{p\in G(c)}\sum_{l=1}^{L_p}\sum_{d=1}^{D}f(l,p)x_{d,c}(l,p).$$

Given the cost models described above, we can now represent the goal of the data cloud via the following integer linear program (ILP).

$$\min_{x,y} \quad \text{OperCost} + \text{ExecCost} + \text{PurchCost} \tag{1}$$

$$\text{subject to} \quad x_{d,c}(l,p) \le y_{p,d}(l)\ \forall c,p,l,d \tag{1a}$$

$$\sum_{l=1}^{L_p}\sum_{d=1}^{D}x_{d,c}(l,p) = 1,\ \forall c, p\in G(c) \tag{1b}$$

$$\sum_{l=1}^{L_p}\sum_{d=1}^{D}x_{d,c}(l,p)q(l,p) \ge w_c(p),\ \forall c, p\in G(c) \tag{1c}$$

$$x_{d,c}(l,p) \ge 0, \forall c,p,l,d \tag{1d}$$

$$y_{p,d}(l) \ge 0,\ \forall p,l,d \tag{1e}$$

$$x_{d,c}(l,p), y_{p,d}(l) \in \{0,1\}, \forall c,p,l,d \tag{1f}$$

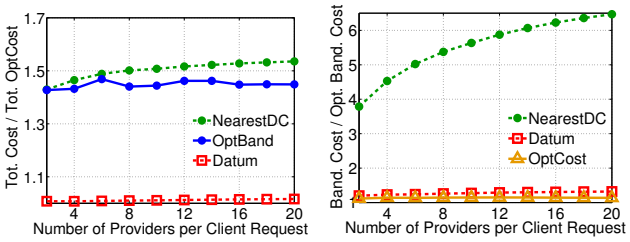
3. Datum

Our first result, stated below, highlights that cost minimization for a data cloud is NP-hard.

THEOREM 1. *The cost minimization problem for a geo-distributed data cloud given in* (1) *is NP-hard.*

More specifically, the reduction leading to Theorem 1 highlights that the data cloud optimization problem is equivalent to the *non-metric* uncapacitated facility location problem Nevertheless, even though our problem can, in general, be viewed as the non-metric uncapacitated facility location, it does have a structure in real-world situations that we can exploit to develop practical algorithms.

In particular, in the case of a data cloud made up of a single data center, though the problem is still an uncapacitated facility location problem, there is a structure that allows us to design an algorithm with polynomial running time that gives an exact optimal solution. The details of the algorithm and corresponding proof can be found in [6].



(a) Total Cost (b) Bandwidth Cost

Figure 1: Datum is near optimal.

Unlike the data cloud cost minimization problem for a single data center, the general data cloud cost minimization is NP-hard. However, the exact solution for single data center case inspires our design, Datum, for cost minimization in a geo-distributed data cloud.

The idea underlying Datum is to, first, optimize data purchasing decisions as if the data market was made up of a single data center (given carefully designed "transformed" costs). Then, second, given the data purchasing decisions, Datum optimizes data placement/replication decisions .

The sketch of Datum is as following. A detailed description of each steps of the algorithm can be found in [6].

Step 1: Define V as the set of all possible subsets of data centers. Reformulate the problem and for data centers, replace subscription d with v to add two new constraints to (1). Those two new constraints guarantee the decoupling of data purchasing and data placement.

Step 2: Aggregate variables x and y with respect to subscription v, and treat the geo-distributed data cloud as a single data center with proper parameter approximation. This leaves the "single data center" problem and thus can be solved optimally in polynomial time.

Step 3: The results of Step 2 determines which quality levels should be purchased and which quality levels should be delivered to each client. Then the remaining problem to determine data placement and data replication an be solved optimally in polynomial time.

4. CASE STUDY

We illustrate the performance of Datum using a case study of a geo-distributed data cloud running in North America. A detailed description of the settings can be found in [6]. Figure 1(a) illustrates the costs savings Datum provides. It highlights that Datum provides near-optimal performance (within 1.6% of optimal) in realistic settings via a polynomial-time algorithm that is provably optimal in the case of a data cloud running on a single data center. Additionally, Datum provides $> 45\%$ improvement over current design proposals for geo-distributed data analytics systems.

References

[1] Microsoft Azure. https://azure.microsoft.com/en-us/, 2015.
[2] Factual. https://www.factual.com/, 2015.
[3] Infochimps. http://www.infochimps.com/, 2015.
[4] Xignite. http://www.xignite.com/, 2015.
[5] The IUPHAR/BPS Guide to Pharmacology. http://www.guidetopharmacology.org/, 2015.
[6] X.Ren, P. London, J. Ziani, and A. Wierman. Joint Data Purchasing and Data Placement in a Geo-Distributed Data Market. http://arxiv.org/abs/1604.02533.