Pricing Approaches for Data Markets

Alexander Muschalle[1], Florian Stahl[2], Alexander Löser[1], and Gottfried Vossen[2]

[1] TU Berlin, FG DIMA, Einsteinufer 17, 10587 Berlin, Germany
forename.surname@tu-berlin.de,
[2] University of Münster, ERCIS, Leonardo-Campus 3,48149 Münster, Germany
forename.surname@ercis.de,

Abstract. Currently, multiple data vendors utilize the cloud-computing paradigm for trading raw data, associated analytical services, and analytic results as a commodity good. We observe that these vendors often move the functionality of data warehouses to cloud-based platforms. On such platforms, vendors provide services for integrating and analyzing data from public and commercial data sources. We present insights from interviews with seven established vendors about their key challenges with regard to pricing strategies in different market situations and derive associated research problems for the business intelligence community.

1 Introduction

The analysis of freely available data, together with commercial and in-house data, is an increasing market segment. One example is market research focused analytics of Web data, with the aid of natural language processing technologies and statistical methods. In order to analyze data sets of such size large IT infrastructures need to be built. However, this is potentially costly as such systems usually have high implementation costs, as well as significant further costs for updating and analyzing data. Even though, we can reduce the problem to a few hundred GB that will fit into main memory, these companies still need capable staff in their IT department who can maintain and program complex in-memory multi-core infrastructures [22]. In particular, for small and medium enterprises (SME) the associated risks with such infrastructures are a strong barrier for innovation.

Arguably, cloud computing could lower this barrier as far as operating the hardware is concerned. It has to be taken into account that, though cloud computing infrastructure might be operated at lower cost, it is still likely to be to expensive for a single SME to rent the hardware necessary to crawl and analyze significant portions of the Web. This is why SMEs are not yet in a position to benefit from the latest research in cloud computing and Web mining. Nevertheless, SMEs often hold unique assets for transferring data into business relevant information, e.g., domain knowledge of a particular niche or relationships to potential customers that are willing to pay for the information.

Only recently, vendors of data, providers of data warehouse solutions, and algorithm providers have started to offer their products as platform-, software-, and data-as-a-service on so called data market places. These data markets supply analysts, business applications, and developers with meaningful data and an eased data access. Moreover, for data producers these marketplaces act as single integration platform. As a result, data marketplaces enable completely new business models with information and analysis tools as electronically tradable goods. The collective storage, analysis, and utilization of data from the Web on a central platform offers many cost savings and high innovation capabilities. Thus, especially for SMEs significant market entry barriers and impediments to innovation are eliminated.

In this paper we present an empirical study with seven data market owners and producers of data associated products and services. In this preliminary study with early adaptors, we collected answers from our interview partners for the following important questions:

1. What are common queries and demands of participants on a data market?
2. Which pricing models utilize beneficiaries for data associated products?
3. Which research challenges for the business intelligence community may arise from the combination of data and data associated services in data markets?

The reminder of this paper is structured as follows: In Section 2 we report on two query categories and seven types of beneficiaries that are common across all interview partners. In Section 3 we discuss the market situation of our interview partners, reveal current pricing strategies and derive major research challenges in Section 4. In Section 5 we discuss related work and conclude in Section 6. Finally in Appendix A we give a brief introduction to our research methodology.

2 What is a Data Market?

In this section we review common elements of business models and business demands of our partners. We start with queries that a data market should be able to answer and then report our observations about seven beneficiaries and their demands.

2.1 Common Query Demands

During our interviews we asked our interview partners for common demands of their customers. We observed a heterogeneous set of queries from which we could derive the following two major scenarios:

Estimate the value of a 'thing', compare the value of 'things'. In the first scenario, customers abstract the data market as a data warehouse that resides on an open set of data sources. These customers utilize the data market for collecting and measuring factual signals from public data sources, such as the Web or 'Open Data' from the UNO, and non-public sources, such as commercial data from Reuters and in-house private data, such as data from an ERP or CRM system. Given the set of available data sources on the data market and the data warehouse-like processing infrastructure, a common user demand is formulating key indicators which describe the value of an immaterial or material good. Finally, the user utilizes these key indicators, together with data sources and the storage and integration platform of the data warehouse, for computing a value of immaterial and material goods and for ranking these goods by this value function. Among many others, our interview partners mentioned the following interesting scenarios:

Ranking influencers on the Social Web. *Which top-10 Web forums influence buying decisions of German BMW motorbike customers? What are the most influential users?* [1]

Value of a Web page for placing advertisements. *On which highly ranked and frequently visited Web pages for 'men's underwear' should I buy advertisement space?* [2]

Value of a starlet or impact of a politician. *Order politicians of the German parliament by the number of mentions in major German newspapers. Does ex-tennis star 'Steffi Graf' reach the right customer set for an advertisement campaign for fast food products?*

Value of a product. *What are top-10 journals for 'my' life science publications? A banker will issue question like: Is the product of a loan requesting company really accepted by their customers? What opinions do their customers state about the product?*

In these scenarios 'fact tables' contain measurable key indicators from publicly available Web data, while dimension tables may contain data from commercial vendors or in house private data, such as product lists. Note, that in this scenario the 'schema' of these dimension tables is formalized through the schema of these commercial or in house data sources. Moreover, all scenarios share the need for aggregating key indicators from the Web and for correlating these key indicators with a monetary transaction, such as decisions on costly marketing measures. Moreover, our interview partners agreed, that their users only require a partial order of the value of objects. Therefore, users are — to a certain extend — willing to tolerate uncertain data samples, such as sentiment and factual data extracted from textual sources.

Show all about a 'thing'. The second class of scenarios is about collecting factual information about a certain thing and integrating this factual information into a universal relation. For decades, this scenario is well known in the information integration community, see [6], but also in the text mining community, see [11] and information retrieval community, see recent work on exploratory search [18], [17]. Our interview partners consider the ability of the data warehouse owner to resolve and reconcile logical objects across a set of heterogenous sources as a major research challenge.

[1] A prominent data collector for answering such queries is the page 'Klout.com'

[2] See also alexa.com and other vendors of products for search engine optimization.

2.2 Beneficiaries and Participants

Throughout our interviews we could identify seven types of beneficiaries that may issue queries against and may profit from the services of a data market. Figure 1 shows the relationship between beneficiaries and services.

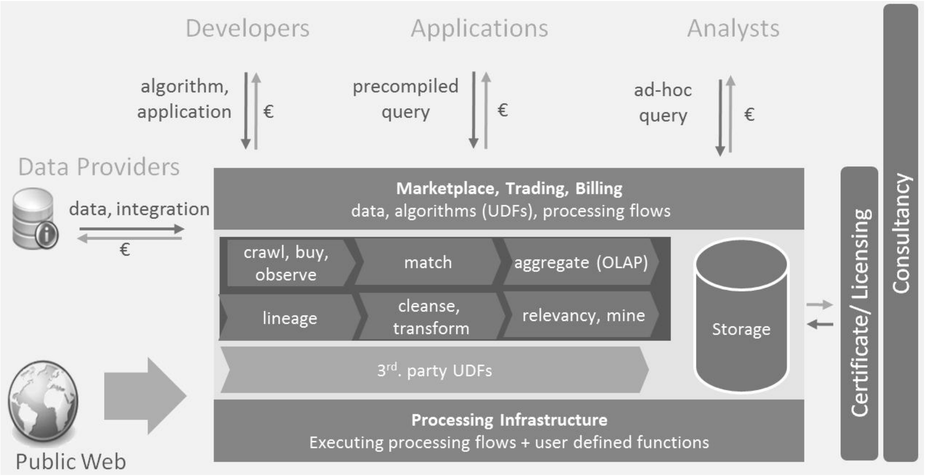

Fig. 1. This figure shows a general schema of a data marketplace for integrating public Web data with other data sources. In analogy to a data warehouse architecture, the schema includes components for data extraction, data transformation and data loading, as well as meta data repositories describing data and algorithms. In additional, the data marketplace offers interfaces for 'uploading' and methods for optimizing, potentially complementary, black box operators with user-defined-functionality, as well as components for trading and billing these black box operators. In return, the 'vendor' of the user-defined-function retrieves a monetary consumption (indicated by the euro symbol) from buyers. Moreover, in the case of large data volumes from the Web, the marketplace relies on a scalable infrastructure for processing and indexing data.

Understanding demands, interests, and needs of different beneficiaries is crucial before we can analyze potential technological challenges. Only then can we start building systems that may solve their problems. In a further step we can think about pricing strategies and prices that these participants are willing to pay. From the interviews we learned about the following seven groups:

1. **Analysts.** Typical members of this group are domain experts, such as M&A experts, sales executives, product managers, brokers, marketing managers and business analysts. The most often used data exploration tools for these experts are Web search engines. To a lesser extend this group utilizes 'traditional' OLAP focused business intelligence tools. Moreover, this group utilizes office products, mostly for aggregating and summarizing results from a data exploration into a meaningful report. From a data market perspective, this group tries to benefit from the sheer endless options of combining

publicly available (Web) data, commercial data (commercial sources) and private (enterprise) data. To do so, this group issues ad-hoc queries against data sources and combines data in a highly interactive way, thereby formalizing knowledge about data sources and data integration steps through their interactions with a data market place. This group is of high relevance for other participants on the data market, since analysts create a demand for frequently requested data and associated data related services through ad-hoc queries.

2. **Application vendors.** Often analyst may not be comfortable in expressing their demands in a formalized machine readable representation. Additionally, the task of data exploration requires many repetitive and labor intensive steps. Application vendors formalize common requirements from analysts into applications. These applications ease data access for a broader group of domain experts. Important examples are business analytics applications [12], mash-up applications [23], customer relationship management applications, and enterprise resource planning applications. These applications formalize common data processing knowledge into a sequence of pre-compiled queries, such as queries for selecting sources or queries for integrating and aggregating data. From a pricing perspective of a data market, pre-compiled query sequences have a high potential to create stable, continuous demands for data and services.
3. **Developers of data associated algorithms.** Both previous beneficiary types, application vendors and analysts, need to integrate data, often from a large and volatile set of sources. Examples are algorithms for data mining, matching, cleansing, relevance computation and lineage tracing. These algorithms vary for different data domains, text data languages, data 'quality' or data 'joinability', and often only fit specific niche domains. Developers may upload these algorithms to a data marketplace as a black box user-defined-function, so other participants of the data marketplace may 'buy' and try out these algorithms. Algorithm developers may buy 'shelf' space and may rent the infrastructure from the owner of the data marketplace.
4. **Data providers.** We learned that our interview partners distinguish between commercial and non-commercial providers of Web data, for example Web search engine owners, such as Bing or Google, Web archive owners, providers of linked data from the Web and Web forum owners who seek commercialization of their forum content. The interviewees also mention government agencies, like the UNO or the World Bank, which provide statistics free of charge. Also, they mention commercial data providers, such as Reuters or Bloomberg among others, which have a long history in the area of selling financial data and geo-data. Data providers utilize market places for storing and advertising data. Some data providers also started offering data integration algorithms.
5. **Consultants.** Our interviewees frequently mentioned consultancy services for supporting analysts in tasks like data source selection, integration, evaluating and product development.

6. **Licensing and certification entities.** On a data market this group assigns a 'branded label' to data, applications and algorithms that may help customers when buying data related products.
7. **Data market owner.** The owner of a data market place is faced with many technical, ethical, legal and economic challenges. Probably the most important economic challenge is establishing a trusted brand and large community. An important technical challenge is developing a common platform for storing, searching and exchanging data and related algorithms, while minimizing algorithm execution time. In particular in Europe, legal aspects may represent significant market entry barriers for trading data across European countries.

In the next section we discuss the different market situations for these beneficiaries. Later (in Section 4), we derive interesting and novel research challenges for the data warehouse and business intelligence community from these market situations.

3 Market Situations and Pricing Strategies

In our interviews we observed that prices for data and associated services mainly depend on demanders' preferences and market structures. This section describes common market structures and associated pricing strategies we observed from our interview partners.

3.1 Current Market Structures

Relevant in the context of this paper are three market structures — *monopoly, oligopoly,* and *strong competition.* In a **monopoly** a supplier holds enough market power to set prices to a level which maximizes profits. In order to do so, suppliers sell fewer quantities at higher price. A monopolist is well advised not to set a single price for a product, but different prices for different demanders (as different demanders have different preferences). This is commonly referred to as price discrimination.

However, these preferences are mostly hidden so that the monopolist tries to motivate demanders to reveal their preferences. This could be done by pricing in respect of quantity, entry level degree, social grouping, etc. In particular in a data market environment the monopolist approach is to set a price, learn from the demanders' reaction by slightly increase/decrease the prices. This enables a monopolist to shape the demand function based on which the supplier is able to maximize profits, by perfect price discrimination.

When monopolists loose their position because one or more competitors enter the market, an **oligopoly** is created, i.e., the market is dominated by a few. This can change the former monopolist's situation dramatically. Effects range from price fights to pooling of interests. Substantial knowledge of the specific industry and complex game theory analysis can potentially help forecasting players' behavior.

Under **strong competition** market prices tend to align with marginal costs. This trend is enhanced especially when marketplaces come up because they naturally promote transparency by enabling demanders to compare products. Thus, individual suppliers lack the market power of setting a profit-maximizing price, but have to face the market price (i.e., selling either for the market given price or do not sell at all). The low marginal costs of data and algorithms can lead to problematic price developments because the overall costs are not relevant for short term decisions. As long as suppliers realize positive gross margins, it is rational to do the transaction at prices near marginal costs. In the long run, this approach is dangerous, because overall costs have also to be covered. As long as suppliers are not able to affect the demanders in a way, that they perceive a unique product quality, suppliers will have to decrease prices. Tweaking overall costs will be the only measure to avoid profit loss. This progress is nothing inherent only to markets of Data as a Service, but are typical in any competitive market with low marginal cost.

We discovered — even among our limited number of interviewees — that the perceived competition differs significantly. Some interviewees were able to mention their *strongest competitors* instantaneously and can thus be categorized as being in a strong competitive environment. Others felt there is no directly competing product or service (*monopolist*). The remaining interview partners knew about similar offerings to their products but were not particular concerned about them. Overall, our interviewees concordantly consider competition as an adequate means to forward welfare and market size growth. Our interviewees consider the market (to date) as big enough and players do not put too much pressure onto each other. Especially in one-to-one competition we found out that there is no willingness to start a price fight.

3.2 Observed Pricing Strategies

It is a common defection, especially in a merely technical domain, to think that prices should be based on costs. In fact, all of our interviewees consider costs only as a limiting factor of reasonable pricing. Rather, they agree that it is almost only the demanders' preferences that determine prices. To what extent suppliers are able to influence a demander's preferences is an issue of marketing operations. A deliberate pricing strategy often turns out to be a major contribution to gain profits rather than any cost reducing measure. According to the interviewees' statements we observed four main categories of pricing models:

1. **Free** data can be obtained from public authorities, such as statistical data[3]. Our interviewees argue that free data can unlikely be sold for money only because it is offered on a market place. Nevertheless, free data available on one's marketplace can help to attract customers which in turn attracts suppliers of commercial data. Moreover, free data can be integrated with in-house or private data and this integration could become valuable.

[3] http://data.gov.uk/

2. **Usage based prices** correspond to the human rationality that each single unit of a commodity raises the total amount of money to pay for. We observed that our interview partners utilize usage based prices for consultancy time or API calls; as an example, in our interview series several partners charged consultancy services per hour. This volume based approach substantially loses power of persuasion if marginal costs converge to zero. Our interview partners expect a trend of falling marginal costs in their product portfolio and expect expect that volume based pricing on markets for data and algorithm may no longer become the first choice among marketplace owners.
3. **Package pricing** refers to a pricing model that offers a customer a certain amount of data or API calls for a fixed fee. For one of our interview partners offering API-based marketing research this was the model of choice. Other vendors, such as Yahoo!BOSS[4] or OpenCalais[5], a Reuters subsidiary, also offer this pricing scheme. Interestingly, none of our interviewees uses APIs together with pure usage based pricing. One reasons could be that these companies sell quantities that are not actually used by the customers or that that accounting overhead is too high. Note that, depending on the package size, API-calls potentially allow the model of arbitrage. Optimizing package based models is a subject of current research, see [14, 15].
4. The **flat fee tariff** is one of the simplest pricing models with minimal transaction costs. It is based on *time* as the only parameter. Among our interview partners, we observed this pricing scheme mainly in regard to software licenses and software hosting. On the one hand, a flat fee tariff provides suppliers and demanders with more safety in planning future activities. This rests upon the fact that time is linear. On the other hand, especially for the demander's side, a flat fee tariff lacks flexibility. A supplier carefully has to bear in mind his product's specific market structure and the demander's preferences. To do so, suppliers could combine a flat fee tariff with flexibility by offering short term contracts.
5. A combination of the previous two is **two-part tariff** pricing. In this scenario customers pay a fixed basic fee and on top of that an additional 'fee per unit consumed'. This pricing scheme is also commonly used by telephone companies. In the data scenario we figured out that one interview partner utilizes the fixed part to cover the fixed costs, where as the variable fee generates the profit. Another example are pricing schemes for software license where prices are calculated by taking a base fee and adding an surcharge depending on the numbers of users who would use the system.
6. **Freemium** is another approach of pricing data and algorithm on marketplaces. The idea is to let users join and use basic services for free and charge them for (premium) services that provide additional value to customers. The payment model for additional services can take any of the forms described above. A couple of our interview partners take exactly that line to ensure a

[4] http://developer.yahoo.com/search/boss/

[5] http://www.opencalais.com/

large attendance to their services. One interview partner realized that there is no value in simply reselling data but rather in offering additional services for data integration . We bear in mind, though, that such a price strategy only works if product's marginal costs are very small, otherwise per unit losings could get out of control.

3.3 The Effect of Product Substitutions on Pricing Strategies

The marginal rate of substitution is an important parameter for pricing strategies in many different market structures. The marginal rate of substitution indicates what quantity of good A is equivalent to a quantity of good B. A perfect substitute (marginal rat of substitution of 1), implies that demanders are completely indifferent wether to buy product A or product B. Under real conditions, perfect substitutes are rare. However, suppliers on marketplaces for data and algorithm have to understand in how far demanders consider the supplier's products or services as substitutable. For example, if a supplier trades a specific tagging algorithm as a monopolist, another supplier with a completely different algorithm, but identical benefit from demander's point of view (tagged text), could gain the whole market, by setting a price slightly below the monopolist's one.

4 Challenges for the Business Intelligence Community

Our interviews resulted in five trends in the area of data markets for the near future. In this section we present for each trend attractive research opportunities for the business intelligence community.

4.1 Growing Number of Data Providers

The first trend is a *growing number of data providers.* One example are Web forum and Web page owners that seek commercializing user-generated-content. Currently, these parties usually sell advertisement space. In the future, these parties will also sell raw and homogenized content or even enrich and aggregate content, thereby allowing for analytical insights. Another example are governmental offices, such as `data.gov.uk` or `www-genesis.destatis.de`. In the past, these organizations mainly offered publicly available reports, often as printouts. Only recently these organizations have started to offer online analytics. Finally, publicly funded initiatives, such as multimillion projects triggered by the European Union or financed by venture capital, will collect and aggregate data and will seek options for commercializing data. Examples are a `www.mia-marktplatz.de`, a market for data of the German Web or `datamarket.com`, a company that mainly collects, integrates, aggregates, and visualizes public available data from governmental institutions.

Challenge 1: Create information extraction systems that can attach structured, *semantically meaningful labels to text data with little human effort. Given that labeled text data, enable an analyst OLAP operations on Web data with the same simplicity as a Web search.*

4.2 Data Markets will offer the Entire Stack for Niche Domains

Recent technological advances, such as sophisticated implementations (e.g., PACT/Nephele [3] implementation of the map/reduce principle [10]), distributed storage and querying architectures (such as HBASE or SenseiDB), or high level batch processing languages (like Pig-Latin[20] or JAQL [4]) drastically simplify the access to even large commodity clusters of data storage and processing nodes. The availability of this technology in the open source community 'potentially' enables each marketplace owner to host the growing number of available data sources. Therefore, *competition and diversification among data markets* will raise, which our interview partners consider as the second important trend. They argue that data markets will provide not only data, but will soon start to offer data associated algorithms and data visualizations. Our interviewees consider to reach different beneficiaries groups as another option for diversification. One example are data market places, such as `infochimps.com` or `factual.com`, which mainly focus on the group of data developers and which provide application programmer interfaces for this group. Our interviewees argue that these market places will soon also serve analysts. Developing such a domain specific stack for niche domains is difficult and requires *tools for creating mash-ups and data supply chains* (like DAMIA[23], KAPOW, or Yahoo!Pipes).

Challenge 2: Enable analysts to create domain specific data processing flows with little costs. Enable and optimize black-box user defined functions in these flows.

4.3 Customers demand data more quickly

From our interviews and market observation we notice a trend towards more brand monitoring. That means that the appearance of brands of consumer goods and also of institutions such as universities are regularly monitored on the Web. This is done towards different ends. On goal is to analyze how a brand is perceived by customers. Another goal is to react to negative comments in order to reduce the harm. In particular the last example falls into the category of publish-subscribe patterns. Companies are only realizing what is possible with Web monitoring and are likely to demand the services in near realtime in the near future. This development is analogous to realtime BI but more complex as the amount of data is significant larger.

Challenge 3: Build systems that can reliably answer brand monitoring queries to indexes, with heavy read and write access, sticking to ACID constraints on a scalable infrastructure in order to enable near realtime brand monitoring.

4.4 Customers will use Product Substitutes

Data markets do not recognize that some products can be substituted by others. That is particulary true for data associated algorithms, such as text mining libraries for extracting factual and sentiment information from user-generated content. As a result, it remains difficult for application developers to identify appropriate data algorithms, in particular for niche data sets. Worse, analysts and developers cannot determine how a particular algorithm is different from competing products. Ideally, customers of data and associated algorithms could try out algorithms on a selected data set before buying. A standardization of data processing mash-ups (see also Challenge 2) would enable product substitution, which in turn would enable competition between suppliers. Generally, competition was seen as very positive by all of our interviewees.

Challenge 4 (example substitution): Given a list of entities and their properties, identify a set of mining algorithms that have been optimized to a very high degree for exactly this data set. Provide a user with an execution sample and recommend an algorithm with regard to execution time, precision and recall.

4.5 Data Markets will offer Price Transparency

On the consumer side the main motivating factor for using data market places is to be able to buy products or services at the right price and in the right quality from a single source. Providing price transparency to users would force data suppliers and algorithm developers to optimize their data sets and algorithms. This may lead to an increase in customers on the market place which again is the most attractive factor for suppliers. Therefore, price transparency can lead to a positive development of the entire market place usage and may increase the revenue for the data market operator. However, for an individual supplier this transparency would eventually cause a drop on sales, since customers of this supplier will eventually substitute the more expensive product with a lower priced product of another supplier.

Challenge 5: Develop incentives for suppliers for price transparency.

4.6 Learn from your Customers

On a data market place, information about data, algorithms and customer behavior is available on a single platform. This information enables a data marketplace to track and derive customer preferences, which are valuable for suppliers. The market place operator could sell this 'secondary' data which would provide another revenue stream. On the economic side, making this information available will bring a data market place closer to a perfect market (i.e., a market on which all information is transparent and available to everyone; one market close to this is the stock market). Moving closer to a perfect market will optimize processes and prices and thus optimize the overall profits and welfare. More than anything else, a broad user base can attract suppliers who offer their data and algorithms on a data marketplace.

Challenge 6: Collect transactional data from customers. Leverage this data for solving Challenges 1-5.

Further challenges. In this paper we focus on pricing strategies. However, we do recognize that that there are more challenges to be addressed, most notably issues of trust and privacy in a cloud-based environment.

5 Related Work

In this section we review publications on pricing of information and data goods in the field of databases as well as in the economic literature.

5.1 Pricing Information Goods

In [2] research areas related to data markets are outlined and a research agenda for the database community is derived. According to the authors, the emergence of data markets bears two challenges regarding the pricing of information or data goods: One major challenge is understanding how the value of data is modified through actors on data markets and which pricing models and services facilitate these data markets. The second challenge is understanding the behavior of market participants and the rules underlying data deals. The authors attribute the first area to the database community and describe the second area rather to the economics community. We, however, believe that this disregards an important basic economic fact: Prices, as soon as a market place is used, converge to a given market price, leaving little scope for price variations.

Most work in this area bases on the assumption that providers are able to set prices in a monopolistic way. Three main strategies exists: flat fee pricing, per use pricing and two-part tariff pricing. Regarding profit maximization the authors of [25] point out that under certain assumptions (monopoly, zero marginal and monitoring costs, homogeneous customers) flat fee pricing and two-part tariff pricing are on par, while *two part tariff* is the most profitable pricing strategy, when consumption is heterogeneous. In [5] it is demonstrated that contingency prices are best if the value of an information good is underestimated, resulting of the uncertainty of demanders as they are not able to experience the quality of the information before buying it.

5.2 Arbitrage, Pricing per Tuple, and Cost Optimizations

Considering arbitrage. The authors of [2] identify two data markets pricing schemes, (a) a subscription based pricing with a query limit and (b) a schema where the price is determined per data set ('tuple'). The same authors discuss four major shortcomings of current pricing schemes: (1) Current price models allow for arbitrage, (2) models base on the assumption that all data sets are of equal value, (3) data acquired once has to be either cached by the customer or paid for again and (4) data providers receive no guidance on how to set their prices. The same authors argue that the first three weaknesses of pricing schemes only occur because of the last shortcoming and propose in [15] an algorithm which validates that a pricing model is free of arbitrage Later, the same authors of [2] also suggest modeling prices in a fine-grained way, i.e., attached to a small data unit (cell, tuple, etc.), and rules about how to transform these prices in course of a query, as this allows for maximum flexibility. They suggest using data provenance to achieve this goal. At the same time, the authors realize that computing theses prices is potentially complex. We argue that this approach does not take into account the fact that data of an arbitrary cell on its own may often have little to no value. If the attached price is equal to the cost of producing the data, one could use the suggested model at least to derive the cost of data and use this as lower bound in price negotiations.

Charging users for optimizations across queries. The authors of [14] investigate how a system could charge users for optimizations across queries. Later, the authors of [24] used a game theory approach where users actively bid for an optimization while the system is value-maximizing and cost-recovering. Such an optimization is enacted if the overall utility (sum of bids minus cost) is maximal. Users are then charged according to their bid. The approach excludes users without a bit from the optimization to give an incentive to reveal true preferences. The authors of [13] suggest a cost estimation model which allows for adaptive shifting from crawling to search and vice versa. While this model estimates the costs of the service, it does not consider the value of the data for customers. In the context of cloud computing the authors of [21] introduce a method to dynamically generate prices upon request. They suggest to use these prices in negotiations or auctions: If a buyer requests data, the seller calculates a price sends it back and buyers may then choose whether to accept or not.

6 Conclusion

The increasing interest in data markets to lever all kinds of available data, public as well as private, in order to create novel consumer and enterprise value is clearly visible. This preliminary study shows seven beneficiaries of data and data associated services. For each of these beneficiaries we discussed potential market situations, pricing approaches and trends. Major research challenges are developing an infrastructure and tools that enable entire enterprises and even individuals to identify insight effectively, from their collection of data assets and from data collections on cloud-based data market places.

Although this study already shows valuable insights, it is limited by the number of participants. In our future work we will continue our study with a broader sample. The disruptions discussed in this study present exciting opportunities for the business intelligence community. We started to address hard problems, the solution of which can greatly impact the future course of data platforms and tools such as identifying text mining services for niche data [7], or investigating data processing infrastructures for large scale data such as [1].

Acknowledgements: We like to thank the following partners in our interviews: Yasan Budak, CFO and owner, Vico Research & Consulting GmbH; Stefan Geissler,CEO and owner, TEMIS Deutschland GmbH; Holger Duewiger, CTO, Neofonie GmbH; Hjalmar Gislarson, CEO and owner, datamarket.com; Ian Mulvany, VP Product Management, mendeley.com; Christoph Tempich, Manager, inovex GmbH and Phillip Mueller, VP Strategy at Thomson Reuters. Alexander Löser receives funding from the German Federal Ministry of Economics and Technology (BMWi) under grant agreement "01MD11014A, 'A cloud-based Marketplace for Information and Analytics on the German Web (MIA)"

References

1. Alexander Alexandrov, Dominic Battré, Stephan Ewen, Max Heimel, Fabian Hueske, Odej Kao, Volker Markl, Erik Nijkamp, and Daniel Warneke. Massively parallel data analysis with pacts on nephele. *PVLDB*, 3(2):1625–1628, 2010.
2. Magdalena Balazinska, Bill Howe, and Dan Suciu. Data markets in the cloud: An opportunity for the database community. *PVLDB*, 4(12):1482–1485, 2011.
3. Dominic Battré, Stephan Ewen, Fabian Hueske, Odej Kao, Volker Markl, and Daniel Warneke. Nephele/pacts: a programming model and execution framework for web-scale analytical processing. In *SoCC*, pages 119–130, 2010.
4. Kevin S. Beyer, Vuk Ercegovac, Rainer Gemulla, Andrey Balmin, Mohamed Y. Eltabakh, Carl-Christian Kanne, Fatma Özcan, and Eugene J. Shekita. Jaql: A scripting language for large scale semistructured data analysis. *PVLDB*, 4(12):1272–1283, 2011.
5. Hemant K Bhargava and Shankar Sundaresan. Contingency pricing for information goods and services under industrywide performance standard. *J. Manage. Inf. Syst.*, 20(2):113–136, October 2003.
6. Jens Bleiholder and Felix Naumann. Data fusion. *ACM Comput. Surv.*, 41(1), 2008.
7. Christoph Boden, Alexander Löser, Christoph Nagel, and Stephan Pieper. Factcrawl: A fact retrieval framework for full-text indices. In *WebDB*, 2011.
8. A. Bryman and E. Bell. *Business Research Methods*. Oxford University Press, 2007.
9. J.M. Corbin and A.L. Strauss. *Basics of qualitative research: techniques and procedures for developing grounded theory*. Sage Publications, Inc., 2008.
10. Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. In *Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.
11. AnHai Doan, Raghu Ramakrishnan, and Shivakumar Vaithyanathan. Managing information extraction: state of the art and research directions. In *SIGMOD Conference*, pages 799–800, 2006.

12. Helena Galhardas, Daniela Florescu, Dennis Shasha, Eric Simon, and Cristian-Augustin Saita. Declarative data cleaning: Language, model, and algorithms. In *VLDB*, pages 371–380, 2001.
13. Panagiotis G. Ipeirotis, Eugene Agichtein, Pranay Jain, and Luis Gravano. To search or to crawl?: towards a query optimizer for text-centric tasks. In *SIGMOD Conference*, pages 265–276, 2006.
14. Verena Kantere, Debabrata Dash, Georgios Gratsias, and Anastasia Ailamaki. Predicting cost amortization for query services. In *SIGMOD Conference*, pages 325–336, 2011.
15. Avanish Kushal, Sharmadha Moorthy, and Vikash Kumar. Pricing for data markets.
16. S. Kvale and S. Brinkmann. *InterViews: Learning the Craft of Qualitative Research Interviewing*. Sage Publications, 2008.
17. Alexander Löser, Sebastian Arnold, and Tillmann Fiehn. The goolap fact retrieval framework. In *Business Intelligence*, volume 96 of *Lecture Notes in Business Information Processing*, pages 84–97. 2012.
18. Gary Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, April 2006.
19. M.D. Myers. *Qualitative Research in Business & Management*. Sage, 2008.
20. Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. Pig latin: a not-so-foreign language for data processing. In *SIGMOD Conference*, pages 1099–1110, 2008.
21. Tim Püschel and Dirk Neumann. Management of cloud infastructures: Policy-based revenue optimization. In Jay F. Nunamaker Jr. and Wendy L. Currie, editors, *ICIS*, page 178. Association for Information Systems, 2009.
22. Antony Rowstron, Dushyanth Narayanan, Austin Donnelly, Greg OShea, and Andrew Douglas. Nobody ever got fired for using hadoop on a cluster. In *HotCDP 2012 - 1st International Workshop on Hot Topics in Cloud Data Processing*, 2012.
23. David E. Simmen, Mehmet Altinel, Volker Markl, Sriram Padmanabhan, and Ashutosh Singh. Damia: data mashups for intranet applications. In *SIGMOD Conference*, 2008.
24. Prasang Upadhyaya, Magdalena Balazinska, and Dan Suciu. How to price shared optimizations in the cloud. *Proc. VLDB Endow.*, 5(6):562–573, February 2012.
25. Shin yi Wu and Rajiv D. Banker. Best pricing strategy for information services. *J. AIS*, 11(6), 2010.

A Appendix: Methodology of Qualitative Interviews

We reported on a study of qualitative interviews with executives of seven European companies that are active in the data market area. In this section we describe our interview methodology.

A.1 Introduction to Interview Techniques

The methodology of this study implements a seven stages approach to interview research as suggested in [16]. The first step is *thematizing the study* followed by *designing the study in consideration of the knowledge to be achieved*. The subsequent steps are *interviewing*, *transcribing*, *analyzing*, *verifying*, and the last step is *reporting/publishing*. Generally it can be distinguished between interviews in quantitative and qualitative research. Whereas quantitative interviews are usually fully structured, qualitative interviews are unstructured or semi-structured [8]. In structured interviews the interviewer has pre-set an interview schedule with questions that are kept simple to allow for easy coding and comparison of results. Thereby, interviewees serve as information delivering subjects. Contrary, qualitative research interviews regard interviewees as participants [16]. Qualitative interviews can be distinguished in almost unstructured and semi-structured interviews. While the first can be regarded as type of conversation, the second follows a guide of predefined open questions or topics. Some authors even encourage to depart from the interview guide and discuss tangents to get as much insights as possible [8].

A.2 Sampling and Interview Process

We selected seven interview partners in executive positions who are involved with developing, consulting, or running data related products and service. In the course of the interviews we extended the list of potential interviews by means of snowballing (i.e., the recommendation of potential new interview partners by interviewees [8]). Our interview partners cover a broad spectrum of the data related products and services, like social media monitoring, text enrichment, consulting or data market places. Often, an interviewee covered more than one of the given areas.

We opted for semi-structured interviews [16] to ensure that all participants underwent a similar interview and to allow for a minimum of comparability. We conducted the interviews via telephone and visualized questions via slides. The main purpose of the slides was ensuring that interviewer and interviewee had a common point of reference. Our interviews covered three topics: First, we asked the interview partners for their current position and experience. Next, we issued questions about their products and business models. Finally, we asked direct questions to examine what an optimal data market place should look like in order to fulfill the needs for the interview partner. The interviews took – on average – 63 minutes.

A.3 Analysis and Verification

In order to preserve the interview we recorded and transcribed the interviews in full [16][8]. For the actual analysis we used a grounded theory approach [9, 19, 8]: First, we identified codes (i.e., themes such as *data has no intrinsic value*); then summarized them in categories (e.g., *factors influencing the price*), next examined relationships between categories and finally drew conclusions from that (e.g., *a market place should offer transparent pricing shemes*). For achieving validity by reaching consent in a conversation or discourse [16] we gave the study results to the interviewees for their approval and seek scientific discourse by publishing this work.