What Next? A Half-Dozen Data Management Research Goals for Big Data and the Cloud

Surajit Chaudhuri
Microsoft Research
surajitc@microsoft.com

→ member of ACM
→ Scientist researching
- Auto-tuning for cloud database systems
- Multi-tenant Cloud Database Systems
- Data transforming / wrangling
- Approximate Query Processing
- Enterprise search over structured Data

ABSTRACT
In this short paper, I describe six data management research challenges relevant for Big Data and the Cloud. Although some of these problems are not new, their importance is amplified by Big Data and Cloud Computing.

Categories and Subject Descriptors
H.2.0 [Database Management]: - General.

General Terms
Algorithms, Performance, Theory.

Keywords
Big Data, Data Analytics, Cloud Infrastructure, Research Challenges.

1. INTRODUCTION
Two accelerating trends are beginning to have impact on the landscape of data management. One is *Big Data*. It is a catch phrase that has many different interpretations but in this paper I will use this term primarily in the context of data analytics. In recent years, here has been a very significant expansion of data analytics [5]. This phenomenon has been fueled by decreasing cost of acquisition of data as more and more data is *born digitally* enabling businesses to collect data that is extremely *fine-grained*. Very low cost of data storage has made it attractive to retain such fine-grained data in the hope of obtaining business insight (e.g., understanding customers). Here are a few specific characteristics of the Big Data phenomenon as they relate to analytics:

- Exploring text and semi-structured data to see if these sources could provide additional insight.
- Narrowing the time gap between data acquisition and acting on a business decision based on the data, sometimes referred to as near real-time business analytics.
- Experimenting with deep analytics beyond the functionality offered by the traditional business intelligence (BI) stack.
- Seeking low cost, highly scalable analytics platforms.

A second disruptive trend that is influencing our field, and indeed the entire computing industry, is the rise of *cloud computing*. The IT infrastructure is gravitating towards a rent model of usage that has the benefit of elasticity (Infrastructure as Service). There is also increasing appetite for web-based services (Software as Service) instead of using packaged on-premise software. The most ambitious aspect of cloud computing is Platform as Service. Unlike Infrastructure as Service, the goal of these platforms is to enable creation of scalable applications without having to think in terms of virtual machines. But, these platforms do impact the model of application development and thus have the danger of lock-in.

These two disruptive trends are shaping our field and it is still not clear what will be the characteristics of the infrastructure and platforms that will emerge from these disruptions. In this paper, instead of speculating on how the industry will evolve, I have tried to identify a few key data management research challenges in the context of big data and the cloud that are hard but where breakthroughs may have significant impact. I will focus on six such problems where I have spent some time working with my colleagues. I have attempted to pick problems with a mix of conceptual, algorithmic, and systems challenges. While not all of these problems are new, the emergence of Big Data and the cloud have amplified the importance of all the six challenges. The rest of the paper consists of brief descriptions of each of these problems, brief remarks on a few other open questions, and the conclusion. The title and the structure of this paper are inspired by Jim Gray's Turing Award lecture.

2. DATA PRIVACY
The term privacy is increasingly being used to refer to all aspects of access to data. With the exploding use of online services and proliferation of mobile devices, the concern about access to and sharing of personal information is growing. Increasing volume and variety of data sets within an enterprise also highlight the need for control of access to such information. While resolution to the policy issues is very important for consumers and enterprises alike, it is also crucial to understand what support for privacy can be provided at the platform level to ease implementation of privacy policies. The three well-known pillars of privacy mechanisms are access control, auditing, and statistical privacy. There are key open questions for each of these facets. For access control, many argue that we should be supporting predicate-based fine-grained control. Such an approach has the potential to provide tighter control on access to information but at the cost of increased complexity of administration as well as performance challenges in providing such support. An example of a mechanism to enable such fine-grained access has been described in [19]. Other experts hold a contrarian view that a rigid access control has been a failure and only coarse-grain access control policies should be implemented [16]. The area of statistical privacy is still

PODS'12, May 21–23, 2012, Scottsdale, Arizona, USA.

at a rather exploratory stage although it has been an active area of research within our field. The initial set of proposals based on data perturbation did not provide the needed combination of high utility along with strong protection from attacks leveraging background knowledge [10]. The model around Differential Privacy [9] provides a sound theoretical foundation. However, whether this framework can be reduced to practice is open to question. Specifically, effective utility of query answers subject to such statistical privacy techniques in real-world SQL applications is unknown. An initial effort to enable such study was the development of PINQ [17]. Probably the most non-controversial area of data privacy and security has been auditing. However, much of today's auditing infrastructure is ad-hoc and unlike statistical privacy and access control, this area has received less attention in research. We also have no conceptual framework to help applications decide how they should deploy a combination of access control and auditing. *Challenge 1: Redefine the abstractions for access control and auditing for data platforms.*

3. APPROXIMATE RESULTS

As the data sets continue to grow larger, the need to do "back of the envelope" calculation to answer queries or progressive refinement of query results is more important than ever. This goal so far has remained unfulfilled except for the simplest of the queries although such functionality is more important for complex business intelligence queries. The difficulty with this problem begins with semantics itself. The simplest semantics could be that of providing a uniform random sample of the result of the query. Yet, such a definition is unsatisfactory. Consider a simple OLAP aggregation query with a measure attribute (e.g., sum of sales by town during winter). The notion of approximation for such a query has two dimensions. One dimension of approximation is how accurate the measure attribute is (e.g., sum of sales for a given town). The other dimension is which of the groups (e.g., towns) are missing in the output. Beyond the semantics, there are significant algorithmic challenges as well. Efficiently obtaining a sample of a simple select-project-join is nontrivial [1][3]. Typically, data is physically organized to serve the access patterns that best match the workload. Not surprisingly, query execution plans are also picked to take advantage of the existing access patterns to reduce cost. The net consequence is execution of even a simple select-project-join query in a database systems often results in the answer stream having ordering properties which prevent using the prefix of the result stream as a random sample of the query. One approach to sidestep this difficulty is to lay out the data randomly and use physical operators for query execution that preserve the randomness of resulting data streams [12][13]. However, the challenge is to ensure that such techniques can be generalized and achieve a significant speed-up over traditional query execution. This remains an open problem and offers opportunities for novel thinking on semantics and alternate execution strategies. *Challenge 2: Devise a querying technique for approximate results that is an order of magnitude faster compared to traditional query execution.*

4. DATA EXPLORATION TO ENABLE DEEP ANALYTICS

One of the big drivers of excitement around Big Data is the expectation that we will be able to identify novel insights in data to drive business decisions. Machine learning is viewed as a key technology that will unlock such insight. Indeed, machine learning has been successfully used for decades in a number of vertical applications (e.g., fraud detection, internet search and advertising). Machine learning toolkits of varying quality and popularity, both commercial and open-source, are widely available. However, effectively leveraging the machine learning toolkits requires understanding of probability and statistics. Even for those who possess that expertise (often called data scientists), the challenge in identifying deep insights from data is quite significant. The critical stage where data scientists lack support in today's infrastructure is in the phase of preparing data for deeper analysis, e.g., in identifying candidate features for machine learning models. Today, for data scientists to be effective they need to be proficient in data querying and use that as the primary means of such exploration. The fundamental difficulties they face to efficiently search for deep insights in data are: (a) How to identify relevant fragments of data easily from a multitude of data sources, (b) How to use data cleaning techniques such as approximate joins across two data sources, (c) How to sample results of a query progressively (see Section 3), and (d) How to obtain rich visualization? While building such a data exploration platform requires systems skills, there are fundamental algorithmic challenges in each of the problems (a)-(d) above as well. *Challenge 3: Build an environment to enable data exploration for deep analytics.*

5. ENTERPRISE DATA ENRICHMENT WITH WEB AND SOCIAL MEDIA

The Big Data phenomenon has provided the opportunity to leverage many diverse sources of data, both structured and unstructured. It is the unique properties of web data such as its vastness, its statistical redundancy and availability of user feedback (via query logs and click information) that has made extraction of structured information (e.g., entities) from web data especially interesting. Such entity extraction techniques have been successfully used for identifying references to specific product names, locations, or people in web pages. For example, the Voice of the Customer class of enterprise applications tries to identify meaningful trends and sentiment information for a given set of products from web pages and blogs. While the rise of these applications represents rich examples of connecting enterprise data to web and social media, building such applications to achieve high precision and good recall is difficult and invariably requires sophisticated custom analytic techniques. Therefore, it would be ideal to identify a set of high precision services that shield the application developers from the above difficulties. For example, given a set of product names and a partial list of their attribute-value pairs, search engine providers and social networking sites could provide a service to identify objects in their respective repositories (web pages, social media posts) that mention one of more of the given entities with high precision and good recall [14][15][23]. Another example of a useful service will be product data conflation based on web information, e.g., discovering common synonyms of products from the web query log and click information [6]. Identifying such high value services, offering a set of derived data assets (e.g., structured contents of infoboxes in Wikipedia), and providing information extraction tools together can help create a platform that has the potential to democratize use of the web and social media data for a much wider class of applications. *Challenge 4: Identify services that given a list of entities and their properties, returns enrichment of entities based on information in web and social media with sufficiently high precision and recall.*

6. QUERY OPTIMIZATION

Query Optimization has been crucial for efficiently answering a large class of complex analytic SQL queries. Even for newer platforms based on MapReduce and its variants, the interest in

leveraging higher level query languages (HiveQL [22], PigLatin [18], and SCOPE [2]) is extremely strong. In such highly parallel platforms, the cost of shuffling data across nodes is considerable and thus query optimization and physical design continue to be critical elements of the infrastructure. It is important to take a fresh look at query optimization because Big Data platforms such as MapReduce introduce changes to some of the fundamental assumptions in query optimization, as explained below. Recall that one of the reasons why MapReduce is a popular framework is because it is easy to express data parallel programs where Map and Reduce functions could be user-defined code. For query optimization, the above is a major departure because optimization of user-defined functions in relational databases was not a central issue. As a consequence, traditional techniques for estimation of sizes of intermediate results need to be revisited. Another related issue is that unlike relational databases, there is no opportunity to create pre-defined statistical summaries of the full data set. Together, these two factors compound the already well-known difficulties of cardinality estimation for query sub-expressions [4]. However, unlike relational systems, MapReduce uses materialization extensively. Therefore, it is attractive to revisit the class of optimization techniques originally pioneered by Teradata to do "optimize and execute" iteratively and inform the next stage of query optimization of properties of intermediate results. Yet another difference is the user expectation that MapReduce systems are batch oriented and hence we can relax the traditional approach to query optimization that had an implicit requirement that optimization time for ad-hoc queries be limited. Although I have cast the above discussion primarily in the context of MapReduce platforms, much of the above is also relevant for parallel SQL database systems. *Challenge 5: Rethink query optimization for data parallel platforms.*

7. PERFORMANCE ISOLATION FOR MULTI-TENANCY

The movement to the cloud is inspired by the opportunity to reduce cost and to leverage the elasticity it offers. For a cloud system provider to deliver on these promises requires multiple users (tenants) to share the same server resources. However, in order for enterprises to feel comfortable to use cloud services, they would also like to achieve performance isolation, i.e., avoid interference with other tenants for the same resources. Although today's cloud service providers offer a service level agreement (SLA) for availability, no such SLA for performance isolation exists. Of course, as soon as multi-tenancy is used, perfect performance isolation is not feasible except at a prohibitive cost to the provider. Therefore, what is needed is a specification of the performance SLAs complete with penalty clauses to mitigate violation of SLAs by providers (as is done today with respect to availability). For our research community, this challenge reduces to defining the framework for multi-tenant data systems. The key technical difficulty inherent in this challenge is that of metering violation of performance SLAs. Therefore, the choice of any model for performance SLAs must also be accompanied by a low-overhead implementation for metering.

Another related problem is the classical challenge of resource allocation among multiple tenants. Even without performance SLAs, this problem still does not have well-founded solutions. For example, the problem of allocation of working memory among adaptive query operators (from multiple queries) in classical relational databases has not received the attention it deserves, despite some recent work [7][21]. In a multi-tenant system, taking into account performance SLAs for doing resource allocation is essential and thus the problem becomes harder. *Challenge 6: Define a model of performance SLAs for multi-tenant data systems that can be metered at low overhead. Develop resource allocation techniques to support multi-tenancy.*

8. Remarks on A Few other Challenges

I have described six of the research problems in Big Data and the Cloud above. However, six is not a magic number and there are indeed several other important issues where our community can be influential. I will briefly mention three such open issues:

Scalable Data Platforms: Recently much attention has focused on this topic [8][20] and so I decided not to discuss this problem in details in this short article. For the foreseeable future, analytics based on relational infrastructure remains essential for the enterprises. Although the MapReduce based infrastructure today lacks much of the maturity of the relational world, it is an emerging ecosystem with much momentum. The rise of this infrastructure offers some unique technical challenges (e.g., see Section 6). However, the longer term goal should be to clearly understand the architectural needs of the spectrum of data analysis platforms for online, near online and batch oriented analysis workloads as each one has unique characteristics.

Operational Business Intelligence: As mentioned in the introduction, there is increasing desire to shorten the gap between data acquisition and business action. For example, a retailer would like to decide on promotions for the next week based on the data collected during this week. For online stores, it is desirable to take action based on data even more quickly. Existing solutions are based on log based shipping, streaming as well as other ETL techniques. However, this field is still at an early phase of its development.

Manageability and Auto-Tuning: One attraction for enterprises transitioning to cloud-based infrastructure is sharply reduced overhead of manageability. Although Infrastructure as Service provides some value in this regard, it is Platform as Service where the providers must fully owe the responsibilities for manageability and tuning of the service. Therefore, the cloud provider needs to develop automated solutions for all aspects of manageability including diagnostics, system parameter tuning, and physical design. On the positive side, with appropriate instrumentation, the provider has the ability to monitor the workload and system events and in fact has the opportunity to tweak such instrumentation much more flexibly compared to packaged software. As a consequence, they are also able to experiment with changes in infrastructure in a seamless manner. Developing such monitoring infrastructure and leveraging deep analytics to support auto-tuning of cloud based services is a very exciting opportunity as well as a significant challenge.

9. CONCLUSION

The increasing interest in Big Data to leverage all sources of available data, public as well as private, to create novel consumer and enterprise value is clearly visible. Our research challenge is to develop the infrastructure and tools that can help enterprises identify signal (insight) effectively from their collection of data assets. We are also witnessing strong movement towards cloud infrastructure. These two disruptions have presented great opportunities to rethink our assumptions. Such significant changes happen rarely. Therefore, as a community, we should seize this opportunity to address hard problems whose solution can greatly impact the future course of data platforms and tools.

10. ACKNOWLEDGMENTS

I am indebted to my talented colleagues in Data Management, Exploration, and Mining Group at Microsoft Research for their insights on the problems described in this paper. Vivek Narasayya has been a great sounding board and a partner in wide ranging brainstorming over the years. Arnd Christian König and Vivek Narasayya read many revisions of this short paper.

11. REFERENCES

[1] Acharya, S., Gibbons, P., Poosala, V., Ramaswamy, S.: Join Synopses for Approximate Query Answering. SIGMOD Conference 1999: 275-286.

[2] Chaiken R. et. al.: SCOPE: easy and efficient parallel processing of massive data sets. PVLDB 1(2), 2008.

[3] Chaudhuri, S., Motwani, R., Narasayya, V..: On Random Sampling over Joins. SIGMOD Conference 1999: 263-274.

[4] Chaudhuri, S.: Query optimizers: time to rethink the contract? SIGMOD Conference 2009: 961-968.

[5] Chaudhuri, S., Dayal, U., Narasayya, V. An Overview of Business Intelligence Technology. Communications of the ACM Vol. 54 No. 8, Pages 88-98.

[6] Cheng T., Lauw H.W., Paparizos S.: Entity Synonyms for Structured Web Search, IEEE Trans. Knowledge and Data Eng., 2011.

[7] Dageville, B., Zait, M. SQL Memory Management in Oracle 9i. In Proceedings of VLDB 2002, Hong Kong, China.

[8] Dean, J., Ghemawat, S.: MapReduce: a flexible data processing tool. Communications of the ACM 53(1): 72-77 (2010).

[9] Dwork, C., Differential Privacy. 33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006), Springer Verlag, Venice, Italy, July 2006.

[10] Dwork,C., McSherry, F., Nissim,K., Smith, A. Calibrating noise to sensitivity in private data analysis. In Proceedings of the 3rd Theory of Cryptography Conference, pages 265–284, 2006.

[11] Gonzalez, H., Halevy, A.Y., Jensen, C.S., Langen,A., Madhavan, J., Shapley, R., Shen, R., Goldberg-Kidon, J.: Google fusion tables: web-centered data management and collaboration. SIGMOD Conference 2010: 1061-1066

[12] Haas, P.J., Hellerstein, J.M.: Ripple Joins for Online Aggregation. SIGMOD Conference 1999: 287-298

[13] Hellerstein, J.M., Haas, P.J., Wang, H.J.: Online Aggregation. SIGMOD Conference 1997: 171-182

[14] Hoffart J. et.al.: Robust Disambiguation of Named Entities in Text, EMNLP 2011.

[15] Kulkarni K., Singh A., Ramakrishnan G., Chakrabarti, S.: Collective Annotation of Wikipedia Entities in Web Text. KDD 2009.

[16] Lampson, B.: Privacy and security - Usable security: how to get it. Communications of the ACM 52(11): 25-27 (2009).

[17] McSherry, F.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. SIGMOD Conference 2009: 19-30.

[18] Olston C. et.al. : Pig Latin: a not-so-Foreign Language for Data Processing. SIGMOD'08.

[19] Oracle Virtual Private Database (VPD). http://www.oracle.com.

[20] Stonebraker, M., Abadi, D.A., DeWitt, D.J., Madden, S., Paulson, E., Pavlo, A., Rasin, A.: MapReduce and parallel DBMSs: friends or foes? Communications of the ACM 53(1): 64-71 (2010).

[21] Storm et al. Adaptive Self-Tuning Memory in IBM DB2. In Proceedings of VLDB 2006, Seoul, Korea.

[22] Thusoo, A. et al. Hive: a Warehousing Solution over a Map-Reduce Framework. PVLDB 2(2), 2009.

[23] Wang C., Chakrabarti K, Cheng T., Chaudhuri S.: Targeted Disambiguation of Ad-hoc, Homogeneous Sets of Named Entities, WWW 2012.