Skolkovo/M.I.T. Collaboration on Big Data

Project Director: Stonebraker

MIT Principal Investigators: Barzilay, Indyk, Jaakkola, Katabi, Madden, Rinard

Russian Principal Investigators: Hirsch (SPbAU), Krouk (ICT), Konushin (MSU), Simakov

1 Centers Focus

Big data is becoming simultaneously the largest opportunity and the key technological bottleneck facing information technology today and into the future. Automated analysis of big data has already transformed both the industrial and scientific landscape. In the industrial front, all modern information technology enterprises (for example, Google, Amazon and Facebook), collect each day terrabytes of data on the behavior and access patterns of their clients. New business initiatives and start-ups are also increasingly being created around data acquisition. For example, oil and gas, healthcare, retailing, and aviation industries are deploying embedded sensors to perform monitoring, asset tracking, and business analytics. Similarly, scientific inquiry today is increasingly data driven, from high-throughput genomic sequencing to remote sensing, astronomy, and oceanography. Petabyte data repositories are becoming common, and exabyte warehouses will soon become a reality. These quantities of data will soon overwhelm communication, processing, and storage systems.

Despite these advances, existing techniques only scratch the surface of what is potentially possible. Indeed, current systems face the data deluge with very simple analysis techniques. While advanced methods are available to extract more useful information from many types of highly valuable data, they do not currently scale to relevant problem sizes. This critical scaling problem and its remedies permeate all stages of the data processing pipeline. For example, the problem of extracting relevant content is a modeling problem. Such models need to be redesigned from the point of view of efficient algorithms and better integrated with databases. Similarly, understanding the content guides how we can efficiently compress and communicate the data in a summary form. Taken together, storage, communication, and processing can no longer be considered in isolation. Indeed, the center's focus is precisely on *enabling transformative advances in big data problems based on integrated solutions*.

Specifically, the center has three key inter-related focus areas:

Business analytics Complex analytics brings big data to life. The ability to run analytics methods efficiently requires a close integration of databases and algorithms. This is currently not the case. For example, common business decisions can be based on identifying groups of users with similar buying patterns or behavior. Such data clustering operations require complex logic beyond simple SQL aggregates used in traditional data bases. Similarly, modeling stock market for trading purposes requires correlating sock performance with covariates over variable trading histories. While such tasks are often simple linear operations, they cannot be expressed as simple SQL aggregates. Solving such problems in todays relational data

base management system (DBMS) is slow and inconvenient. We propose a collaborative effort to integrate complex analytics (modeling) with core database design.

Wireless networking Much of data generation and consumption will be over wireless networks. For example, data from embedded sensors monitoring oil or gas fields, implanted medical devices, and various sensors are all disseminated over wireless networks. Similarly, mobile video traffic is expected to increase close to 100 fold in the coming few years. However, this data deluge is rapidly overwhelming existing qqWiFi and cellular networks. Major advances in wireless technologies are needed to be able to support content transfers at the expected levels. This is a multi-faceted effort. It involves content modeling for effective compression, and communication protocols that exploit such modeling. We bring together expertise in algorithms and networking to realize this potential.

Unstructured data A great deal of the most valuable data today is unstructured and user generated. This includes content such as text, videos, and images that are distributed by the users on the web or social networks. There is tremendous potential in exploiting this data for business analytics, creating new services, or advertising. However, there's a real concern across the business technology landscape that the efforts and solutions are substantially lacking at this point. Similarly, from the point of view of users, they are unable to effortlessly explore these vast data sources. This is a cross-cutting effort. We need advances in machine learning, computer vision, and natural language processing to interpret and prepare content at the required scales. We need communication protocols and database designs that are well-integrated to support content modeling.

2 Center's Technical Thrusts

2.1 Innovations in Data Base Management Systems: SciDB

We propose to work with Pavel Vehilkov, Roman Simakov, Artyom Smirnov, and Konstinin Knizhnik, (Pavel: affiliations of these folks????) on SciDB and its applications in Russia. Two specific use cases drive this collaboration. The LYRA astronomy project (specifically Oleg Bartunov of the Sternberg Astronomical Institute of Lomonosov, Moscow State University) is proposing to use SciDB. Lyra will collect 200-400 TB of raw data which needs to be cooked into observations and analyzed in order to develop a unified and precise star catalog of 300 million stars. LYRA requires features not currently in SciDB such as spherical geometry and nested arrays. We propose to work with the Russian team above to extend SciDB in these directions and to implement LYRA on SciDB. Another project where SciDB plays a crucial role is a project with Sberbank, a leading Russian financial institution, where we are working with (Pavel: who exactly?) on a big data analytics platform for the bank. Our solution will include real-time calculation of risks for large volume of transactions during the day and Monte Carlo simulation techniques applied to market behavior.

2.2 Innovations in Interpreting Unstructured Data

Unstructured data require substantial processing to be useful. For example, queries to social networking sites lead to diverse matches in terms of blogs, video, or pages. The results need to be communicated in a summary form to be useful, especially for mobile devices. The same problem, in reverse, appears in ingesting diverse sensor feeds over bandwidth limited channels where summarization is used to effectively compress and model correlated feeds. Such tasks are made possible with machine learning, natural language processing, and computer vision techniques that properly interpret diverse data, produce effective summaries for delivery or ingest, and anticipate next steps. The ability to predict next steps is, in particular, a great enabler. For example, likely user actions can be used for targeted advertising or likely changes in sensor feeds may

be used to avoid costly anomalies. Predictive modeling is also effective in improving communication by guiding pre-fetching and data compression.

Major advances are needed to permit accurate content modeling at the desired scale. Algorithms have to run (sub-)linearly in order to be relevant and this already precludes basic operations such as many types of clustering. Indeed, the focus is on the emerging interface between efficient algorithms and content modeling. Examples include approximation algorithms that decompose modeling problems so as to achieve much greater parallelization, or methods for restructuring modeling operations as efficient "on the fly" computations with data streams.

Another major focus is on enabling predictive modeling with privacy guarantees, especially for mobile users. Content modeling typically relies on pooling data so as to leverage the common experience of many users. We seek to characterize the resulting accuracy - privacy trade-offs, and develop scalable distributed modeling approaches that operate within prescribed policies.

2.3 Innovations in Data Transmission and Wireless Technologies

The third focus of the center is on issues surrounding data ingest. The explosion of wireless sensors that will, over time, track everything of material significant on the planet is going to compound this issue. We expect to work on techniques to economize wireless bandwidth in these situations. These solutions include advanced interference management techniques, better coding and transmission schemes that increase spectral efficiency (bits/s/Hz), scalable protocols for heterogeneous radio networks, and novel protocol and software architectures, particularly cross-layer approaches that work across the traditional network layers. On this aspect of our research, we propose to work with (Dina: who exactly?) In addition, power consumption is a major problem with such sensor devices, and we expect to explore means to economize power consumption. Finally, security and privacy considerations will gate the acceptance of wireless technology in applications such as insurance telematics. Hence, we works on improving the security of embedded wireless systems and enhancing the privacy of users of mobile applications. Our approach incorporates physical layer techniques, modern cryptographic protocols, and system software. This work will be in conjunction with (Dina: who exactly?)

3 Educational Contributions

Our educational plan is two-pronged. The first stage consists of activities that can be accomplished while SkTech and its is still being formed. These include:

- Summer schools: We will organize intensive, week-long advanced courses on topics relevant to our project, including algorithms, machine learning, databases, wireless communication, natural language processing, and computer vision. We will build on our extensive experience in organizing such courses. E.g., PI Indyk organized several summer schools at Massive Data Algorithms (MADALGO) Center in Aarhus University, Denmark; PI Katabi at MIT Professional Education Program; Alexander Kulikov and and Andrew Goldberg organized Microsoft Data Structures and Algorithms School in Saint Petersburgh; Alexander Kulikov also organized NoNA Summer School on Complexity Theory.

  The activities will be coordinated with Yandex School on Data Analysis as well as Computer Science Club in Saint Petersburg. The courses will be targeted at graduate students and postdocs, and will provide the background necessary to conduct research and educational activities in the respective areas.

- Online courses: We will also facilitate undergraduate courses in computer science through MITx, MIT?s new online learning initiative led by Prof. Anant Agarwal at CSAIL. The initiative will provide the infrastructure for teaching large undergraduate courses over the Internet.

In the second stage, we will build on the accomplishments and expertise gained during the first stage, to establish permanent courses taught at SkTech and possibly other universities in Russia.

4 The Team

Big data is inherently a core computer science problem. Our team combines scientists from the top Computer Science department (MIT EECS, Computer Science Area) and Computer Science Laboratory (CSAIL) in the United States with the very best Russian computer scientists. Together, the team members bring all of the expertise required to fully exploit the opportunities that big data offers.

PI Stonebraker has a proven track record of successful collaboration with Russian scientists, specifically in the SciDB project (a big data project in the area of scientific computing). PIs Stonebraker and Madden lead the recently established Intel Big Data center at MIT. PI Katabi leads the recently established Wireless Center at MIT (sponsors include XXX). Our team's participation in these centers will set up a synergistic interaction between these centers and the Skolkovo project. For example, the centers can help establish connections between Russian scientists and students and the international companies that sponsor the centers. The centers can also host Russian scientists and students on visits to MIT.

The MIT PIs have an established track record of producing intellectual property, then building successful companies around that intellectual property. Together, the PIs hold XXX patents issued within the last five years, with many more under submission. The MIT PIs also have a proven track record of innovation that builds on their intellectual property, having founded multiple successful startups. Examples include XXX. A team led by two of the MIT PIs also won the Web/IT track of the MIT 100K Entrepreneurship Competition.