The Beckman Report on Database Research Summary

Big Data Management Systems 2023 - Exercise 1

Iliadis Viktoras 8180026

Big Data has become a defining challenge of our time due to the affordable generation and processing of data, as well as the democratization of data access. However, the current approach to managing data faces significant disruptions to accommodate the challenges brought by the era of Big Data.

The first challenge is to be found in the need for scalable systems that will manage increasingly larger and faster data sets. The distributed computing field has achieved success in scaling up data processing for less structured data on large numbers of unreliable, commodity machines through the use of constrained programming models such as MapReduce. However, as declarative languages become increasingly adopted in the field, stronger and more affordable query optimizers, and set-oriented query execution engines to scale large clusters of manycore processors must be developed. Also, researchers should explore ways of leveraging specialized processors to process large datasets, which will push for a reconsideration of parallel and distributed query-processing algorithms.

The research community must also get accustomed to a range of emerging storage technologies that are developed for different purposes. Furthermore, there is a need for new scalable techniques for high-speed data stream processing and new algorithms that are optimized for hardware behavior. For data processed only once, it is not cost-effective to store and index it in a database. It should be stored as a binary file and interpreted as a structured record when read later. NoSQL systems were developed to handle high rates of data capture and updates for schema-less data, but they provide weak guarantees. A new class of big data systems emerged to provide full-fledged database-like features. This presents an opportunity to revisit programming models and mechanisms for data consistency and develop new techniques for robust applications.

The data-driven world requires a diverse range of data management systems to handle the various data types, shapes, and sizes that exist. Cross-platform integration is necessary and diverse programming abstractions are required. Platforms that can handle both raw and cooked data are required, with lazy computation being beneficial. Big data systems should become more interoperable like ”Lego bricks,” and cluster resource managers provide inspiration at the systems level. Workflow systems and tools for managing scientific workflows are also useful in handling data processing workflows.

Research efforts should focus more on end-to-end processing of data, with tools that can go from raw data to extracted knowledge with minimal human intervention. Multiple tools are needed, each solving a step of the pipeline, and they should be customizable and easy to use. Open-source tools are needed to benefit from ongoing contributions by the research community. Explanation, provenance, filtering, summarization, and visualization requirements are critical for making analytic tools easy to use, and knowledge bases are essential for improving accuracy and finding domain experts. There is a trend towards creating and using community-maintained knowledge centres for data analysis, but more work is needed on tools to help groups of users collaboratively build, maintain, query, and share domain-specific knowledge bases.

From a data platform standpoint a PaaS would be optimal, however to realize the complete vision there are several challenges to be found. These challenges include achieving elasticity, managing data replication, automating system administration and tuning, ensuring multitenancy, supporting data sharing, and integrating with hybrid clouds. These challenges require resolving issues such as network latency, availability, consistency, performance, programmability, and cost. The cloud enables data sharing at an unprecedented scale, but it also raises new challenges such as finding useful public data, distributing costs, and protecting data if the current cloud provider fails. Moreover, hybrid clouds require interoperation among database services across the cloud, on-premise servers, and mobile devices, posing real-time and mission-critical data management challenges.

The database research community needs to address the change brought about by Big Data by managing not just the data, but the people involved as well. This includes considering human factors related to query understanding and refinement, identifying relevant and trustworthy information sources, defining and refining the data processing pipeline, and visualizing relevant patterns.

In addition to research challenges, big data have created new and exacerbated already existing community challenges. Database education is outdated and needs to change to reflect current technological advances. Additionally, the demand for data scientists is increasing due to big data. They need training on a broad range of skills and interdisciplinary programs that

will allow computer science to synergize with other fields. Finally, there is concern over the research culture of the field. There needs to be a shift towards valuing fewer publications per researcher per time unit and emphasizing large systems projects, end-to-end tool sets, and data sharing to pursue the big data agenda effectively.

The rise of big data presents exciting new research challenges related to processing, handling, and exploiting data. Approaches to education, involvement with data consumers, and research evaluation and funding need to be rethought. It is time to adapt to the new challenges presented by big data.