Big Data: From Querying to Transaction Processing

Karthik Ramachandra
Computer Science and Engg. Department
IIT Bombay
karthiksr@cse.iitb.ac.in

S Sudarshan
Computer Science and Engg. Department
IIT Bombay
sudarsha@cse.iitb.ac.in

ABSTRACT

The term Big Data has been used and abused extensively in the past few years, and means different things to different people. A commonly used notion says Big Data is about "volume" (of data), "velocity" (rate at which data is inserted/updated) and "variety" (of data types). In this tutorial, we use the term Big Data to refer to any data processing need that requires a high degree of parallelism. In other words, we focus primarily on the "volume" and "velocity" aspects.

As part of this tutorial, we will cover some aspects of Big Data management, in particular scalable storage, scalable query processing, and scalable transaction processing.

This is an introductory tutorial for those who are not familiar with the areas that we will be covering. The focus will be conceptual; it is not meant as a tutorial on how to use any specific system.

1. OVERVIEW

This tutorial is organized in three parts. In the first part, we cover basics of distributed data storage, including distributed file systems such as GFS/HDFS, fragmentation and replication of relational data, and scalable key-value stores such as BigTable/HBase.

In the second part, we cover query processing, starting from parallel processing of relational operations, and moving on to the map-reduce framework and programming model, and frameworks such as Hive, which integrate relational operations with map-reduce, and provide high level query languages.

In the last part of the tutorial, we focus on scalable transaction processing and consistency issues, including the Brewer CAP theorem, ACID vs BASE, and issues of eventual consistency. We also briefly outline the key ideas underlying some recent highly scalable transaction processing systems. We conclude the tutorial with a discussion on open problems and challenges, and research directions.

2. TIMELINE

The tutorial will run for 2.5 hours. Each of the three parts will take about 45 minutes, with 15 minutes for questions and discussion.

3. BIOGRAPHIES

S. Sudarshan completed his Ph.D. at the Univ. of Wisconsin, Madison, in 1992. He was a Member of the Technical Staff in the database research group at AT&T Bell Laboratories, from 1992 to 1995, and since then he has been at the Indian Institute of Technology (IIT), Bombay, where he currently holds the post of Institute Chair Professor in the Computer Science and Engineering Department. Sudarshan's research interests center on database systems, and his current research interests include holistic optimization spanning the programming language/database boundary, query optimization for big data, testing of database applications, and keyword queries on semi-structured data. He has published widely on these and other areas in leading international conferences and journals. He is currently an associate editor of ACM TODS, IEEE Trans. on Data Engineering, and IEEE Data Engineering Bulletin. He is a co-author of a database textbook, Database System Concepts, 6th Ed., by Silberschatz, Korth and Sudarshan, which is widely used across the world.

Karthik Ramachandra is currently pursuing his Ph.D. in Database systems at IIT Bombay. His current research is in the area of Holistic Optimization of Database Applications, which is an interdisciplinary area between database query optimization, program analysis and optimizing compilers. His work has been awarded the Microsoft Research India PhD Fellowship and the Yahoo! Key Scientific Challenges Award. During his PhD, he interned at the Microsoft Jim Gray Systems lab, where he built a prototype for a query processing engine for big data. Prior to starting his Ph.D., he has worked as a senior consultant and an application developer at ThoughtWorks Inc. for 5 years, where he has designed and developed enterprise web applications. He has built opensource tools in Java, Ruby, Erlang and other functional programming languages. He has presented talks, demos and tutorials about these tools and technologies at various industry forums.

The 19th International Conference on Management of Data (COMAD),
19th-21st Dec, 2013 at Ahmedabad, India.