

Big Data Analytics 2013



Invited talk

SCOPE: Parallel Databases Meet MapReduce

Jingren Zhou, *Microsoft Corporation*

In this talk, I describe a cloud-scale distributed computation system, called SCOPE, targeted for massive data analysis over tens of thousands of machines at Microsoft Bing. SCOPE combines benefits from both traditional parallel databases and MapReduce execution engines to allow easy programmability and deliver massive scalability and high performance through advanced optimization. Similar to parallel databases, the system has a SQL-like declarative scripting language with no explicit parallelism, while being amenable to efficient parallel execution on large clusters. An optimizer is responsible for converting scripts into efficient execution plans for the distributed computation engine. A physical execution plan consists of a directed acyclic graph (DAG) of vertices. Execution of the plan is orchestrated by a job manager that schedules execution on available machines and provides fault tolerance and recovery, much like MapReduce systems. SCOPE is being used daily for a variety of data analysis and data mining applications over tens of thousands of machines at Microsoft, powering Bing and other online services.