Workload Management for Big Data Analytics

Ashraf Aboulnaga [#], Shivnath Babu [*]

[#]*Cheriton School of Computer Science, University of Waterloo*
Waterloo, Ontario, Canada
ashraf@uwaterloo.ca

[*]*Department of Computer Science, Duke University*
Durham, NC, USA
shivnath@cs.duke.edu

I. Introduction

Parallel database systems and MapReduce systems (most notably Hadoop) are essential components of today's infrastructure for Big Data analytics. These systems process multiple concurrent workloads consisting of complex user requests, where each request is associated with an (explicit or implicit) service level objective. For example, the workload of a particular user or application may have a higher priority than other workloads. Or a particular workload may have strict deadlines for the completion of its requests.

The research area of Workload Management focuses on ensuring that the system meets the service level objectives of various requests while at the same time minimizing the resources required to achieve this goal. At a high level, workload management can be viewed as looking beyond the performance of an individual request to the performance of an entire workload consisting of multiple requests.

Questions addressed by workload management research and technologies include: How to implement different priorities for different workloads? How to isolate the performance of one workload from the effect of other workloads? What is the best way to do request scheduling and admission control? What are good mechanisms and policies to control the allocation of resources to workloads statically and dynamically? How to define a workload and associated requests within that workload? How to monitor request performance, resource consumption, and data access patterns to ensure that workload management is effectively achieving its goals? How to ensure that workload management goals are met even in the presence of failures?

This tutorial will discuss the fundamentals of workload management, and present tools and techniques for workload management in parallel databases and MapReduce. Workload management for parallel databases is an established topic, and most parallel database systems have sophisticated workload management tools. The tutorial will present some of these tools as case studies and discuss the underlying techniques that they use. Workload management for MapReduce is still a fledgling research area, and the tutorial will discuss recent advances in this area and future research directions.

II. Tutorial Outline

- Introduction
  - Problems caused by lack of workload management
  - How are workloads defined?
  - How are goals of workload management defined?
  - Actions that a workload manager can take to achieve its goals
- Workload isolation
  - Effects of workload interference
  - How to define workload isolation?
  - Mechanisms for workload isolation
- Resource allocation
  - Static vs. dynamic resource allocation
  - Techniques for resource allocation in parallel databases
  - Techniques for resource allocation in MapReduce systems
- Scheduling
  - Techniques for scheduling in parallel databases
  - Techniques for scheduling in MapReduce systems
- Control at admission time
  - Need for admission control
  - Techniques for admission control
  - Role of performance modeling
- Current trends and wrapup
  - Cloud computing: resource elasticity, heterogeneity, pay-as-you-go costing
  - "No one size fits all" approach to system design and deployment
  - Wrapup