Book Review

Data Management in Machine Learning Systems
Matthias Boehm, Arun Kumar and Jun Yang
Synthesis Lectures on Data Management
ISB: 9781681734965 (Paperback)
9781681734972 (ebook) 9781681734989 (hardcover)
Morgan & Claypool Publishers
www.morganclaypool.com

In the last few years Machine Learning applications become widespread as it infuses intelligence in various system activities. Realizing this fact the authors have brought this synthesis with nine chapters. In the introduction chapter they have given an overview of preliminaries with emphasis on machine learning life cycle and its users. Besides they have outlined the scope of this book.

In the second chapter on ML Through Database Queries and UDFs, they have presented a view of hoe MS performs using database system. Linear Algebra is basically used in ML algorithms and hence a useful discussion on it with fundamentals is initiated in this unit. In the third chapter on Multi-table ML and Deep Systems Integration, the authors the joins are explained as it is a prerequisite for ML. Integrating linear algebra into ML techniques using RDBMS is required to understand the database management in ML. This is outlined comprehensively in this chapter.

In the fourth chapter on Rewrites and Optimization, they stressed the fact that the ML programs are optimized using rewrites. The optimization systems with Rewrites constitute this part. The optical and physical rewrites are described further.

In the next unit on Execution Strategies, the authors specified that three execution strategies underlie the large scale ML systems. Various executions such as data-parallel, task-parallel and hybrid executions are illustrated to gain good understanding. In the sixth chapter on Data Access Methods, there is a survey of the currently used techniques for data access such as caching and buffer pool management and compression types are stated.

In the seventh unit on Resource Heterogeneity and Elasticity, the two challenges that exist in data management in ML, the Resource Heterogeneity and Resources Elasticity are described. The eighth chapter, Systems for ML Lifecycle Tasks the core issues of preparing data for ML is presented. This includes the process of sourcing, model selection and management and deployment of ML models into production are presented with a good review of many existing studies.
The last chapter of Conclusions includes the issue of integrating ML into data streams which is concerned with data-driven applications, data intensive workload characteristics and data system support for ML. This chapter is followed by a lengthy and descriptive bibliography of many related works.

This volume is a very useful work for the students and researchers on ML who can well understand ML infusion in data management.

Hathairat Ketmaneechairat
King Mongkut's University of Technology
North Bangkok. Thailand