Concepts and Models of Parallel and Data-centric Programming

MapReduce Design Patterns – Introduction

Lecture, Summer 2020

Simon Schwitanski
Dr. Christian Terboven

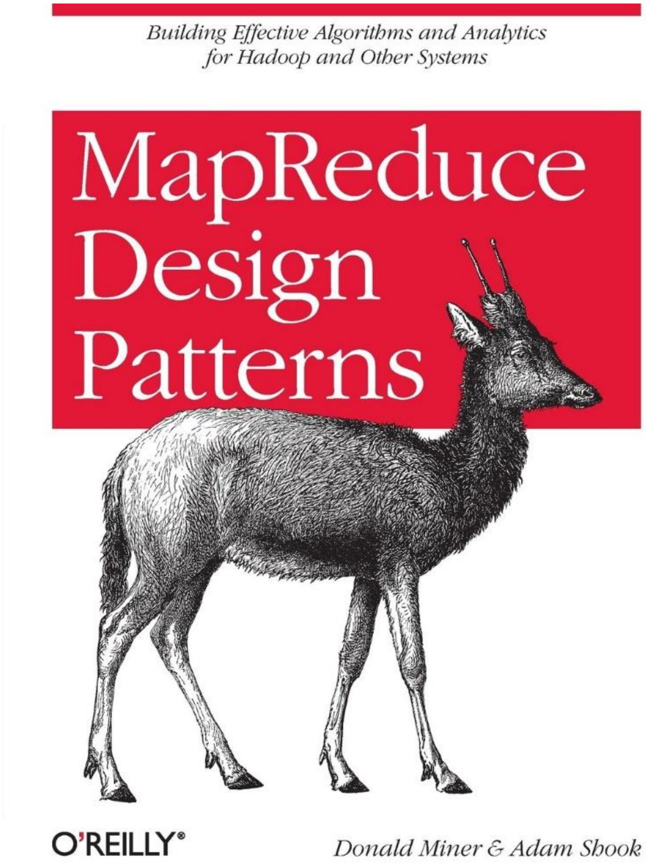
Outline

0. Organization
1. Foundations
2. Shared Memory
3. GPU Programming
4. Bulk-Synchronous Parallelism
5. Message Passing
6. Distributed Shared Memory
7. Parallel Algorithms
8. Parallel I/O
9. **MapReduce**
10. Apache Spark

a. MapReduce Programming Model
b. Parallelizing MapReduce
c. Hadoop Ecosystem
d. Hadoop Distributed File System
e. Yet Another Resource Negotiator
f. Comparison to Other Approaches
g. MapReduce Design Patterns
    a. Summarization Patterns
    b. Filtering Patterns
    c. Data Organization Patterns

Literature

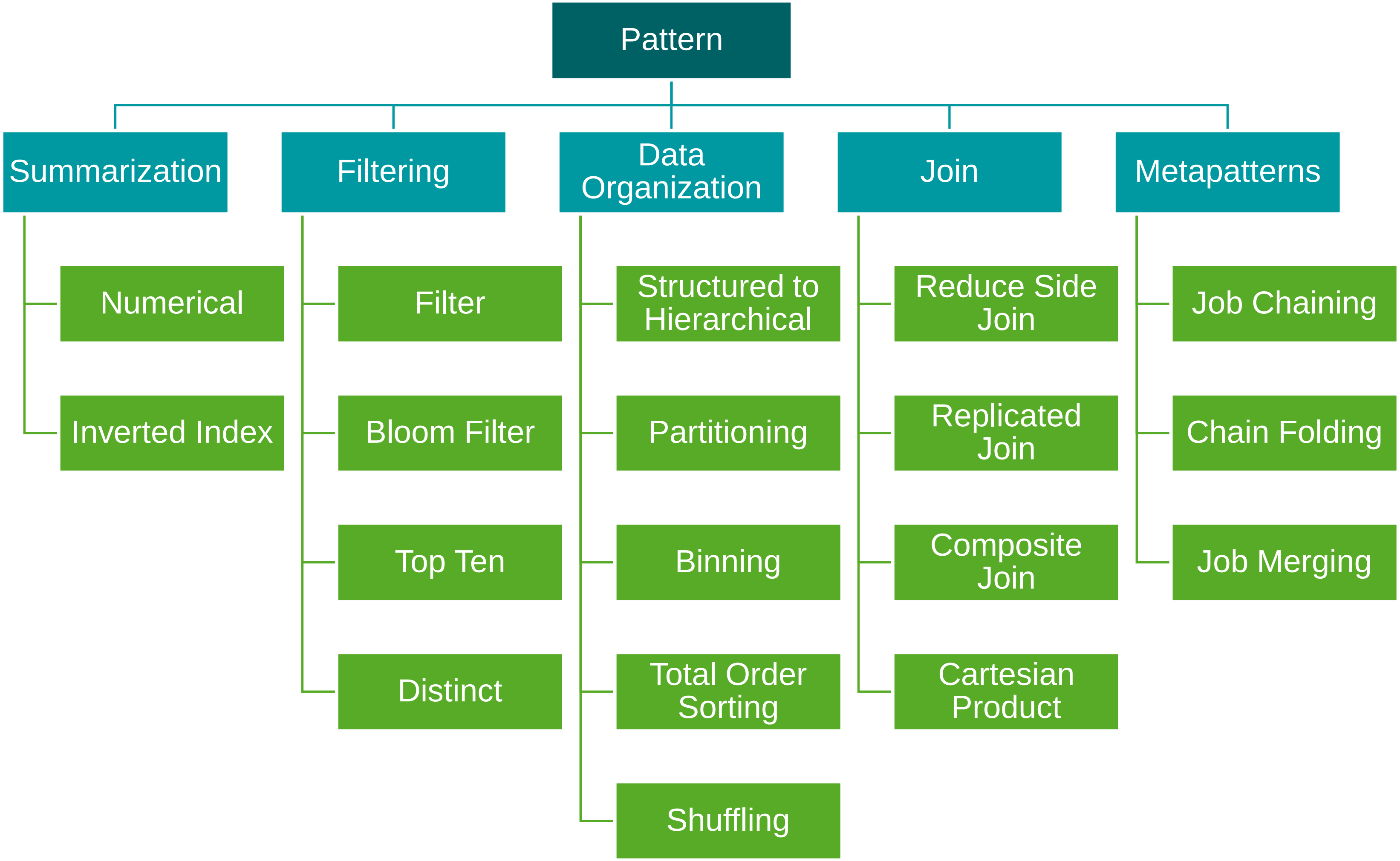
- Miner, Donald and Shook, Adam. “MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems”, O'Reilly Media, 2012
- Slides based on this book
- Note: Reading this book is not required, neither for the lecture and exercises nor for the exam.


Building Effective Algorithms and Analytics for Hadoop and Other Systems
MapReduce Design Patterns
O'REILLY
Donald Miner & Adam Shook

What is a Design Pattern?

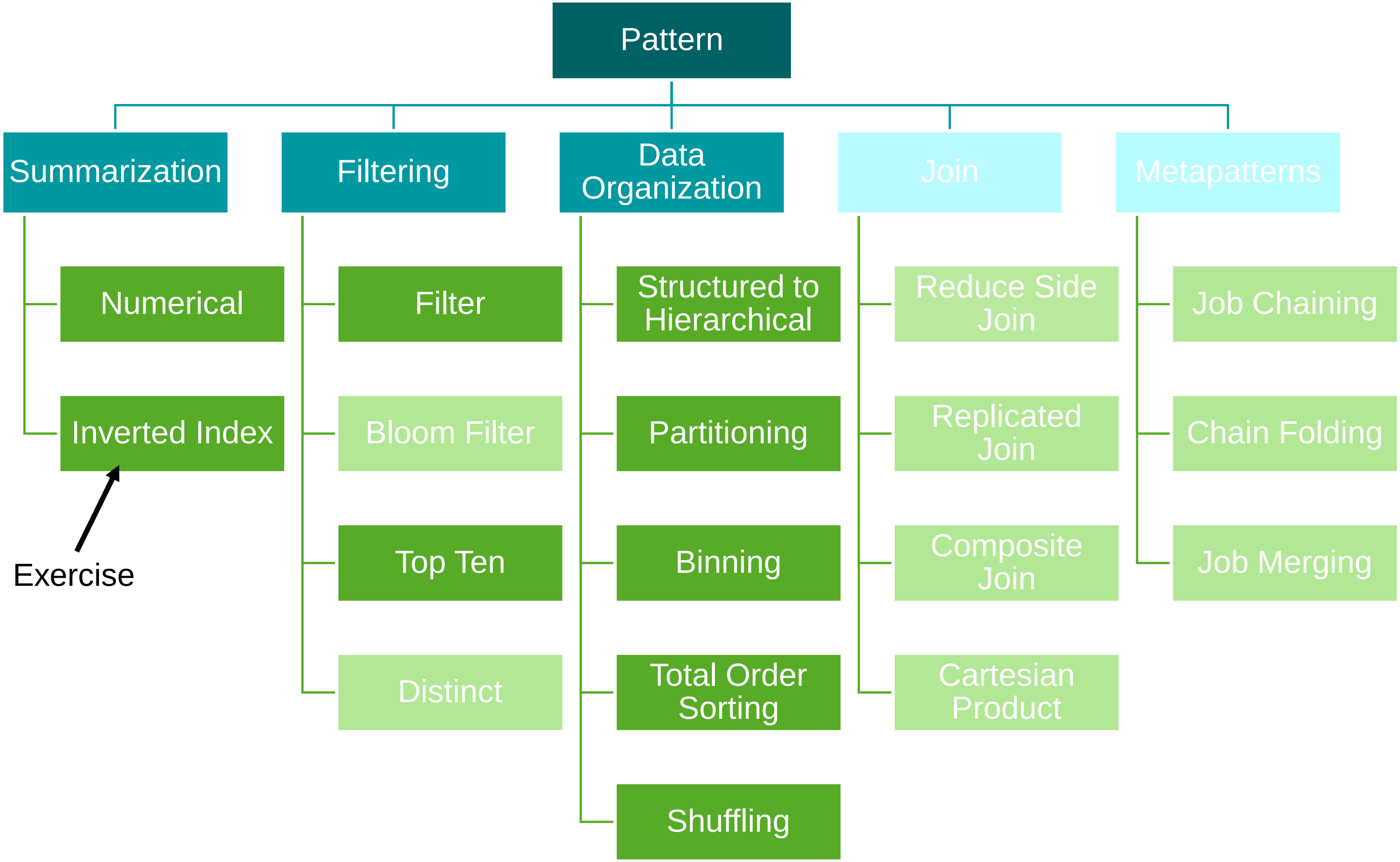
- Blueprints for solving common problems in reusable and general way
- Popular SE design patterns in “Gang of Four” book
- Simplifies discussion with other developers and reading of other developer’s code
- MapReduce design patterns: Smaller problem space than SE patterns
- However: Design patterns incorporate knowledge of experienced developers
  - Helpful for new MapReduce programmers

Pattern Landscape



- Pattern
  - Summarization
    - Numerical
    - Inverted Index
  - Filtering
    - Filter
    - Bloom Filter
    - Top Ten
    - Distinct
  - Data Organization
    - Structured to Hierarchical
    - Partitioning
    - Binning
    - Total Order Sorting
    - Shuffling
  - Join
    - Reduce Side Join
    - Replicated Join
    - Composite Join
    - Cartesian Product
  - Metapatterns
    - Job Chaining
    - Chain Folding
    - Job Merging

Pattern Landscape


Pattern
Summarization
Filtering
Data Organization
Join
Metapatterns
Numerical
Inverted Index
Filter
Bloom Filter
Top Ten
Distinct
Structured to Hierarchical
Partitioning
Binning
Total Order Sorting
Shuffling
Reduce Side Join
Replicated Join
Composite Join
Cartesian Product
Job Chaining
Chain Folding
Job Merging
Exercise