Contents

Contents

Contents

Contents

Contents

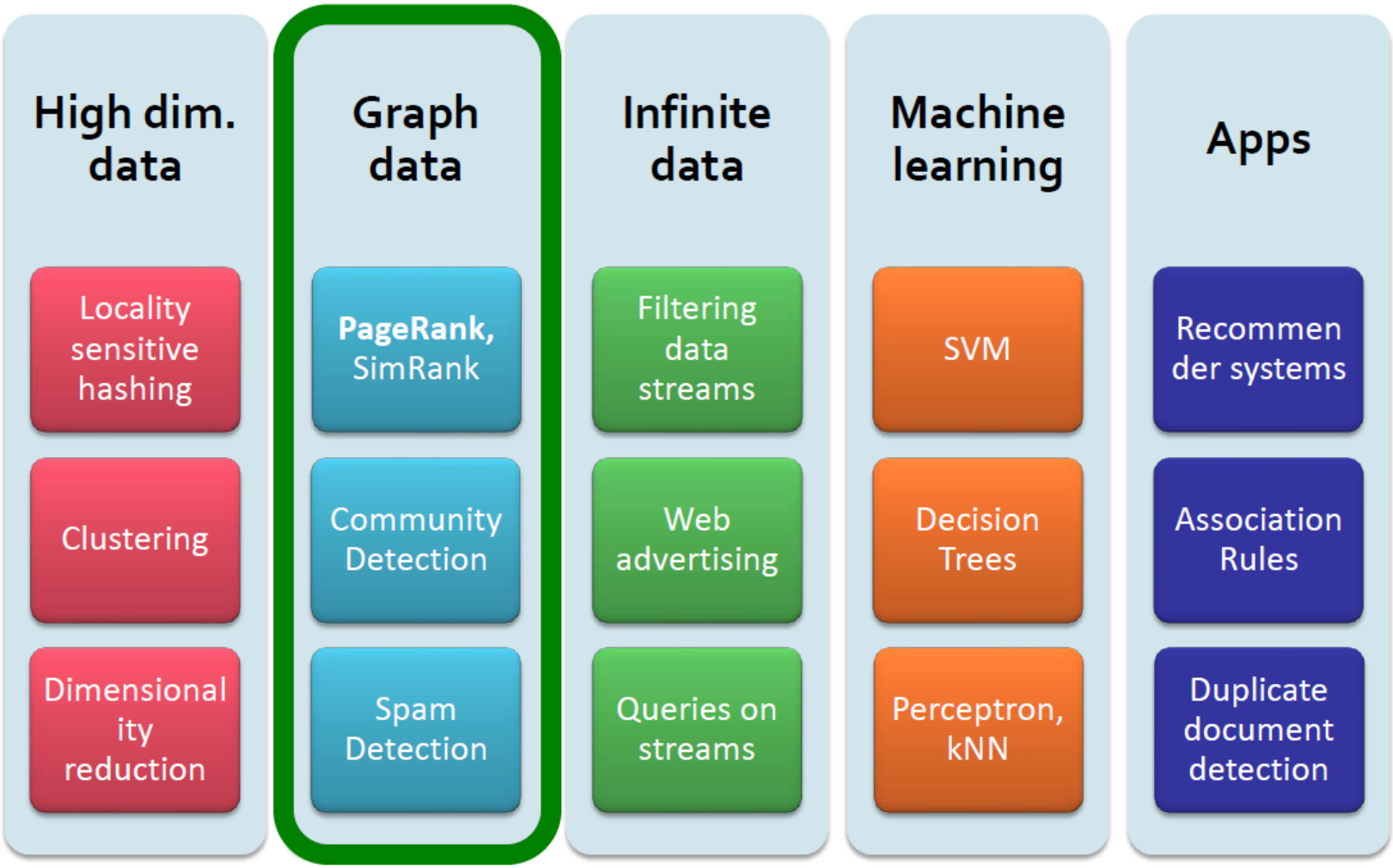
Outline

Content

- For this last course we want to present other topics more related to Big Data issues :
- New types of algorithms on graph data, or e-business data
- Introduce Spark as the state-of-the art framework for ML at scale
- Slides from several MOOCs, advanced materials
  - Mining Massive Datasets book site
  - Spark MOOC from Berkeley (several course available)
  - www.coursera.org/learn/recommender-systems (very detailed)

Outline

New Topic: Graph Data!



Outline

- Invented and deployed by Google (Larry Page co-founder) ;
- Algorithm on graph data (see also second part of the course) ;
- Improve search of web documents, but cannot be directly applied to enterprise search ;
- Idea : simulation of a random surfer walk. Book chapter 5 p.154+.

Page Rank formula

- If M is the transition matrix, to avoid dead-end and spider traps, the iterative formula is modified from $V' = MV$ to $V' = \beta MV + (1 - \beta)\frac{e}{n}$
- Where β in the range $0.8 - 0.9$ and n number of nodes, e vector with all components equal to one ;
- Math background : Stochastic matrix, Markov process, Perron-Frobenius thm.
- Recall general context of search engine process, the steps are :
  1. Crawl the web and build an inverted index ;
  2. Handling the search request ;
  3. Present the search results to the user.

Outline

- Invented in the 90's on e-business context and content website with huge catalog
- Unsupervised ML
- From simple computation , Matrix decomposition to complex, state-of-the-art framework like this year Amazon open source dsstne (destiny)
  https ://github.com/amznlabs/amazon-dsstne

Contents

Contents

Contents

Contents

Contents

Outline

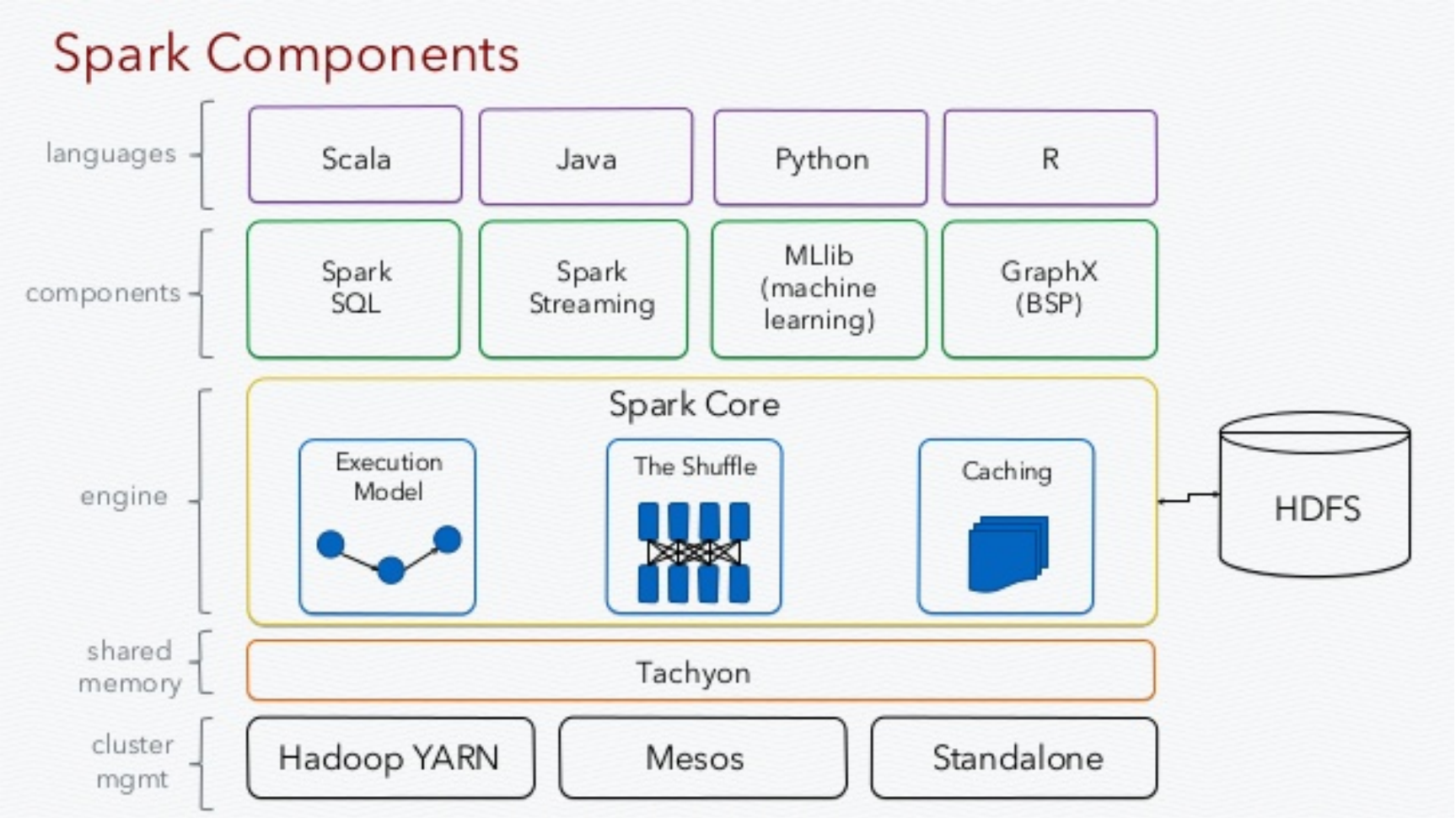
- Overcome the map-reduce V1 limitations : lot of reading/writing on disks ;
- In-memory computation ;
- Rewriting of Mahout into MLLib ;
- Main pillars of Spark ;
- Ecosystem...

Main Components

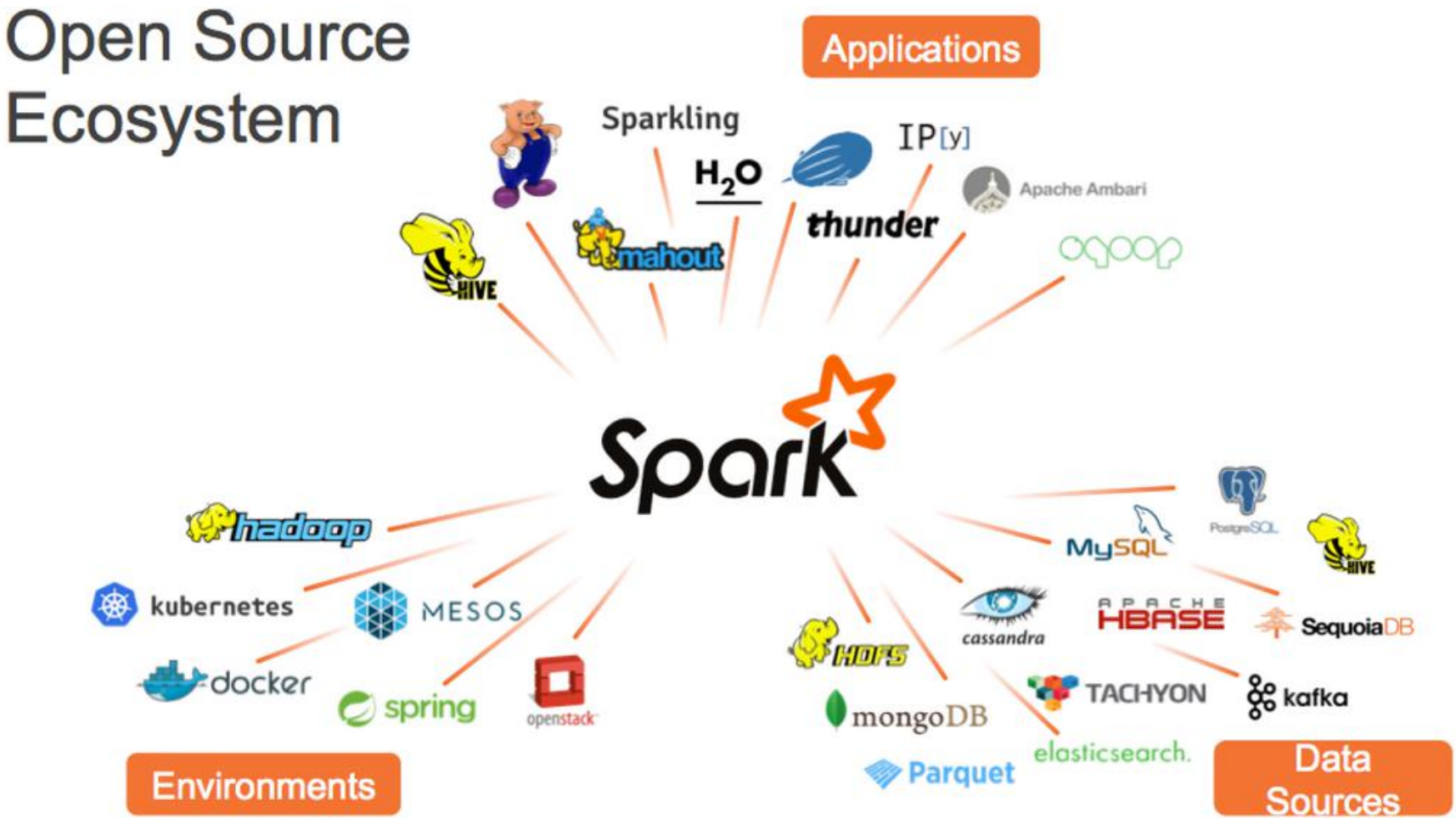
Distributed computation engine designed for big data and in-memory processing

- Interactive and batch analytics
- Up to 100x faster than Hadoop
- 5-10x less code than Hadoop
- Efficiency and scalability
- Fault-tolerance

Main Components with Layers



Spark Ecosystem



Contents

Contents

Contents

Contents

Contents

Machine Learning + Big Data context

Review books materials