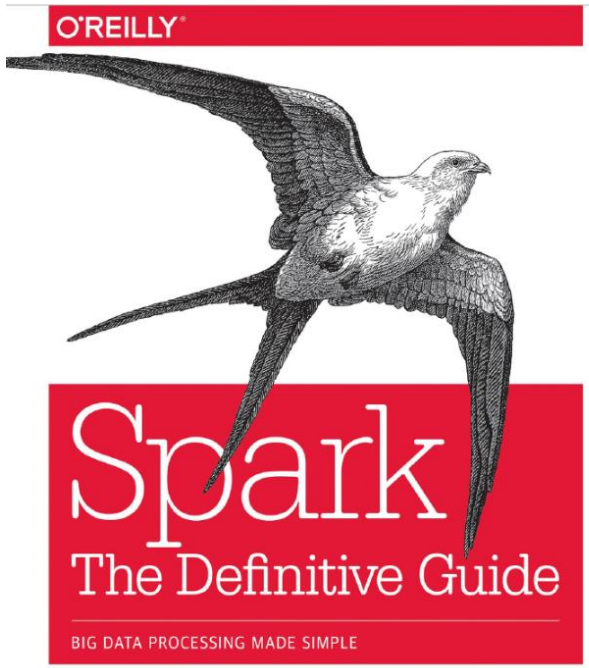
Introdução ao Apache Spark

A filosofia e o propósito do framework



Capítulos Abordados

1. What is Spark?

Filosofia do Hadoop:

Armazenar e processar grandes volumes de dados em vários computadores commodities.

Componentes Básicos:

<Frameworks in Hadoop Ecosystem/>

Batch Processing
Low level Java API

Lightning-fast unified analytics engine

Specialized systems (ML, Streaming, Graph, SQL, etc).

<Spark vs MapReduce/>

	
Fácil de programar	Difícil de programar
Pode-se realizar processamento Batch, Streaming e Machine Learning tudo na mesma aplicação	Batch
Possui modo interativo	Não possui modo interativo (Exceto por frameworks como Pig e Hive)
Escrito em Scala	Escrito em Java
Armazena dados em memoria (e disco quando necessário)	Armazena dados apenas em disco
Processamento em memória, podendo ser sem utilização de escrita e leitura em disco rígido	Processamento em disco - com utilização de escrita e leitura em disco rígido

<Finally: what is Spark?/>

Apache Spark is a unified analytics engine for large-scale data processing.

Unified Analytics Engine	Funcionalidades	Batch ETL, Data Analytics, Machine Learning, Streaming
	Linguagens	Scala, Python, Java, R, SQL (Em breve: .Net).
	Unifica diversos contextos	Por exemplo: SQL + Machine Learning + Streaming
Computing engine for large-scale data processing	Processamento em memória	Muito mais rápido que o MapReduce
	Acessa diversas fontes de dados	Azure Storage, Amazon S3, GCS, Cassandra, Kafka, Bancos de dados relacionais, MongoDB, etc.
	Utilizável em diversos ambientes	No seu próprio notebook, Standalone, YARN, Mesos. Obs: Não necessita de um HDFS!
Libraries	Padrões	SQL, MLlib, Spark Streaming, GraphX
	Terceiras	https://spark-packages.org/