How do I learn Apache Spark?

14 Answers

Suman Bharadwaj, was big data developer at Intel

19k Views

Start learning core concept of Spark - Resilient Distributed Dataset (RDD) . For further information regarding RDDs, refer to the below links. It contains lot of information on how RDD works.

1. https://www.usenix.org/system/fi...
2. http://www.cs.berkeley.edu/~mate...

Additionally, you can also watch this fantastic video on RDD by Matei Zaharia:

1. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing

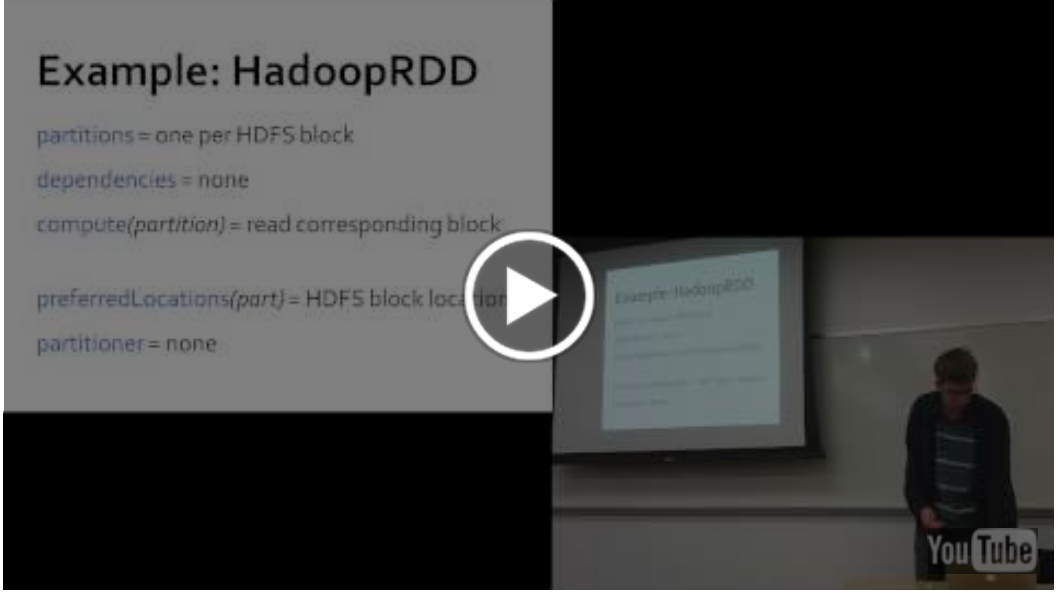
To help you get started with Spark, screen-casts are available on the SPARK website. Play around with Spark for some time to get your self familiar with the spark shell

1. First Steps with Spark - Screencast #1
2. Spark Documentation Overview - Screencast #2
3. Transformations and Caching - Spark Screencast #3
4. A Standalone Job in Scala - Spark Screencast #4

Although, Spark provides APIs for multiple languages, learning Scala will definitely help. A small introduction to Scala shell is available-

1. Introduction to the Scala Shell

Once you are comfortable with above concepts and played with Spark for some time. Get into the internals of Spark. Matei Zaharia talks about the internals of SPARK in the below video. [An excellent video, it helped me a lot]



Now it's time to learn optimization techniques:

1. Tuning Spark - Spark 1.0.0 Documentation
2. A powerful Big Data trio: Spark, Parquet and Avro

Also, you can register to the Spark channel in Youtube. Spark summit, meetup and other videos are available in the Youtube Channel Apache Spark

Other links which helped me are :

1. Differences at shuffle side between Hadoop and Spark- Page on berkeley.edu
2. Spark with Java 8 - Making Spark Easier to Use in Java with Java 8

Spark is fantastic and easy to learn. You'll be amazed to see what 2-3 lines of Spark code can do.

Happy coding and welcome to lightning fast analytics !

Updated 14d ago • View Upvotes

More Answers Below. **Related Questions**

How do I learn Hadoop, MapReduce and Apache spark programming online with hands on in coding?

Where/how can I have hands on experience with Apache Spark?

Is Apache Spark tough to learn?

I'm new to Spark and is interested in learning more. Are there any good blogs on spark other than the documentation provided in the website?

What prerequisites are needed to learn Apache Spark? Is Java knowledge compulsory?

Vik Paruchuri, Founder at dataquest.io

2k Views • Vik is a Most Viewed Writer in Apache Spark.

I've found that not being able to setup Spark easily is a major barrier to being able to learn it. At Dataquest, we recently created a free course where you can interactively learn Spark by running real code in your browser. You can get started here: Introduction to Spark.

The edX course Introduction to Big Data with Apache Spark also looks interesting.

Written 22 Sep • View Upvotes

Bạch Giang, Data Analyst

8.6k Views

Scala basic
Scala School
A Tour of Scala: Case Classes

Spark basic
4. Working with Key-Value Pairs
Data Exploration Using Spark
Page on latrobe.edu.au

Python lambda, reduce, filter
Python Tutorial: Lambda Operator, filter, reduce and map

Simple Spark SQL - Try normal SQL and see magical things happen
Spark SQL Programming Guide
Spark 1.0.2 ScalaDoc
Note that Spark SQL currently uses a very basic SQL parser. Users that want a more complete dialect of SQL should look at the HiveQL support provided by HiveContext.

Nice hack with Spark
Why Apache Spark is a Crossover Hit for Data Scientists
Data Exploration Using BlinkDB

Start with Python
Page on ipython.org

Some Python APIs equivalent to Scala
PySpark

MLLib
Linear Methods - MLlib - Spark 1.0.2 Documentation

Recommender
Building a food recommendation engine with Spark / MLlib and Play
Crouching Data, Hidden Markov
All-pairs similarity via DIMSUM | Twitter Blogs

Spark + Mesos
mesos/spark
How to build Apache Mesos on Mac
How to set up mesos for running spark on standalone OS/X
Installing Mesos onto a Mac with Homebrew
Cluster & Framework Mgmt
Running Spark on Mesos
Cloudera + Mesos + MapReduce + Spark + Chronos - is it possible? Are there similar alternatives?
6 Tutorials for Apache Mesos: Hadoop, Spark, Chronos, and More

Spark + CHD
Running Spark Applications

Updated 6 Sep, 2014 • View Upvotes

Arush Kharbanda, Big Data Architect, sigmoid.com
4.2k Views • Arush is a Most Viewed Writer in Apache Spark.

Spark Provides a really Good documentation. You can follow these steps.

1. Go through the overview. Spark Overview - Spark 1.2.1 Documentation
2. Get started with Spark - Quick Start - Spark 1.2.1 Documentation
3. Go through the programming guide Spark Programming Guide 's (All of them, along with other documents - COnfiguration, cluster modes, security, etc.)
4. Then you can look at the Spark Youtube channel, to learn more.
5. If you want to use Spark with the Scala API. I would suggest Programming Scala and Functional Programming in Scala

You can do some hand's on as you go through the material and try out stuff as you go through it.

Written 16 Mar • View Upvotes

Ranga Raghunathan • Request Bio
3.2k Views

Some good resources mentioned here.
I found this book to be very useful w.r.t applying Spark for analysis. Also provides a decent delving into Spark.

Advanced Analytics with Spark: Patterns for Learning from Data at Scale: Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills: 9781491912768: Amazon.com: Books

Hope it helps !

Written 20 May • View Upvotes

Related Questions

What exercises do you recommend for learning Apache Spark and MLlib?

What are good books or websites for learning Apache Spark and Scala?

How do Apache Spark Streaming and Apache Samza compare to each other?

What is the future of Apache Spark?

What are some interesting applications of Apache Spark?

What's the best practice to run scheduled jobs like crontab with Apache Spark? (especially for PySpark)

I have been assigned to a new product team doing machine learning stuff within Transaction Risk Management Services at Amazon for an undergrad...

What are some useful resources to learn Apache Spark in Java?

Do I need to learn Hadoop first to learn Apache Spark?

Why is Apache Spark called a general-purpose graph execution engine for Hadoop?

Is there any tool in Google similar to Apache Spark in functionality?

What is the architecture of Apache Spark SQL? How do the Spark SQL queries run on SchemaRDD's background?

What are kinds of applications is Apache Spark not suitable for?

What are the main differences while using Apache Spark over Hadoop?

How do I contribute code to Apache Spark?