Developing Spark Applications with Python & Cloudera

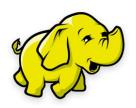
WHY SPARK WITH PYTHON & CLOUDERA?

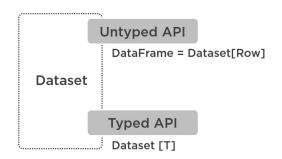


Xavier Morera
HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

@xmorera www.xaviermorera.com







WHY SPARK WITH PYTHON & CLOUDERA?





But Why Apache Spark?



Why Should Use Apache Spark?



Speed

But Why Apache Spark?



Logistic Regression: Hadoop vs. Spark

Apache Spark

??

MapReduce



Execution time: Lower is better



Logistic Regression: Hadoop vs. Spark

Apache Spark

0.9 s

MapReduce

1110s

Execution time: Lower is better



Ease of Use

But Why Apache Spark?



```
demos = sc.textFile("/user/cloudera/spark-course/")
demos.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a + b)
```

Word Count Example

Many lines in Hadoop

A couple of lines in Spark

Easy to learn and full of features



Unified Engine for Big Data

But Why Apache Spark?



Spark SQL

Spark Streaming

MLlib

GraphX

sparkpackages.org

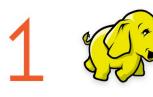




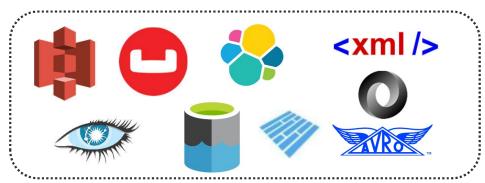








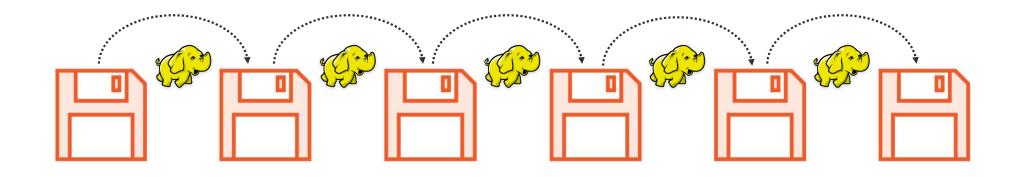




iterative algorithms + interactive data mining tools

Motivation behind Spark, as stated in the original paper







History of "What Led to Spark" in Big Data

Beginning of time

Humans need to store and analyze data

1880's

Modern automated computation
US Census Bureau

1991

Internet is born Need for parallel computing

2006

Hadoop is born

17 Century

Statistical analysis is born Early warning system for Bubonic plague

Before 1990's

Computing era begins, programming languages, data storage, data centers, databases and business intelligence

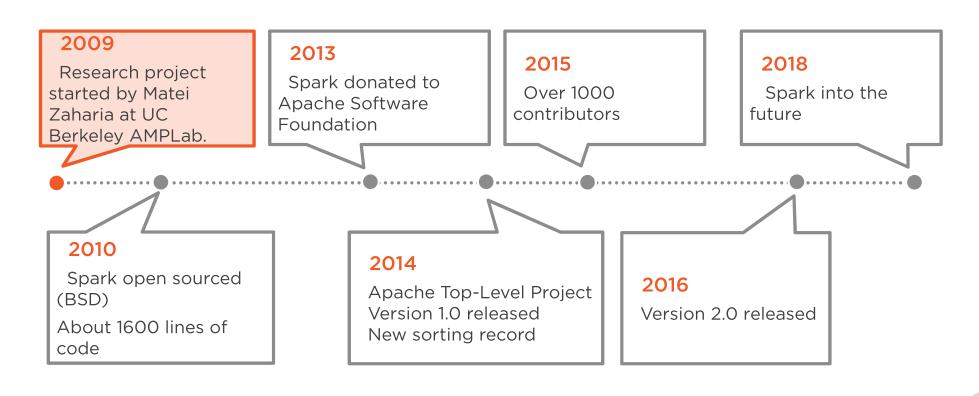
2003 & 2004

MapReduce & GFS papers published by Google





History of Apache Spark



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



Matei Zaharia, Mosharaf Chowdhery, Yaftagata Dao, Ankur Dave, Joetin Mis, Marphy McCauley, Michael J. Frankin, Scott Shenior, Ion Stoica

Abstract

We prove the probability frame (SEII), c due. The province trap districts the first properties of the probability of the probability of the probability of the properties of t

I introduction

Chair companie have two less haplantes [19] and project [19] [19] in the state of t

Example in pathway consists the fore-developquarties from materia for some application that quarties cons. For example, Fregel [22] is a system throater graph computation that large intermedian in

per persolvers, e.g., tolet o me had revel delectives to memory and on abbie queter acres diese. In the paper, so propose a ser deviate has called in about Austriani desent (SCO) that articles reliane dies rener as a benef many of applications. SCOs on faultiferrati, pushful data emotions that be soon on placify provid extremelate results in emoting, control

aspinite from using a club to at Grapestone. The case is deletery in designing EUC to a deliting a gauge-state gas in designing EUC to a deliting a gauge-state gas interests that one provide facility between the contraction of the contraction between the superposed spaties in a matched case (e.g., selfs in a whele). With this involves, the ends very term of the contraction beard on the superposed spaties in the first contraction beard on the superposed spaties in the data contraction beard on the superposed spaties in the state of the contraction of the superposed spaties are not under the first contraction of the superposed spaties are stated as a superposed spaties and the superposed spaties are stated as the superposed spaties are stated as the superposed spaties are stated as the latest ten attention, whose bandwidth is to be born than the other ten attention, where bandwidth is to be born than the other ten attention of the superposed spaties are contracted as the provincing.

date in mr. The altern item to obtained provide half tolerance by legging the interdesentions could be believed a date in the legging the function with a date if it points and date. It is possible to the standard date. It is possible to of an EEO in lost, the EEO has recognite international to the legging of the legging of the entire that here it was determined and the EEO in a recognite international to the legging of the

"The ignoracy for data in come 1970, any for system often a farmagnification processing, belongs, and to design for the data of the 2012

http://tiny.bigdatainc.org/rddpaper

Index of /releases

Name

Last modified Size Description

[ICO]

[PARENTDI	R] Parent Directory
[TXT]	spark-release-0-3.html 2017-10-18 05:57 11K
[TXT]	spark-release-0-5-0.html 2017-10-18 05:57 10K
[TXT]	spark-release-0-5-1.html 2017-10-18 05:57 9.6K
[TXT]	spark-release-0-5-2.html 2017-10-18 05:57 8.1K
[TXT]	spark-release-0-6-0.html 2017-10-18 05:57 13K
[TXT]	spark-release-0-6-1.html 2017-10-18 05:57 9.3K
[TXT]	spark-release-0-6-2.html 2017-10-18 05:57 9.7K
[TXT]	spark-release-0-7-0.html 2017-10-18 05:57 15K
[TXT]	spark-release-0-7-2.html 2017-10-18 05:57 10K
[TXT]	spark-release-0-7-3.html 2017-10-18 05:57 10K
[TXT]	spark-release-0-8-0.html 2017-10-18 05:57 20K
[TXT]	spark-release-0-8-1.html 2017-10-18 05:57 15K
[TXT]	spark-release-0-9-0.html 2017-10-18 05:57 22K
[TXT]	spark-release-0-9-1.html 2017-10-18 05:57 16K
[TXT]	spark-release-0-9-2.html 2017-10-18 05:57 12K
[TXT]	spark-release-1-0-0.html 2017-10-18 05:57 23K
[TXT]	spark-release-1-0-1.html 2017-10-18 05:57 16K
[TXT]	spark-release-1-0-2.html 2017-10-18 05:57 13K
[TXT]	spark-release-1-1-0.html 2017-10-18 05:57 25K
[TXT]	spark-release-1-1-1.html 2017-10-18 05:57 14K
[TXT]	spark-release-1-2-0.html 2017-10-18 05:57 29K
[TXT]	spark-release-1-2-1.html 2017-10-18 05:57 14K
[TXT]	spark-release-1-2-2.html 2017-10-18 05:57 12K
[TXT]	spark-release-1-3-0.html 2017-10-18 05:57 28K
[TXT]	spark-release-1-3-1.html 2017-10-18 05:57 14K
[TXT]	spark-release-1-4-0.html 2017-10-18 05:57 33K
[TXT]	spark-release-1-4-1 html 2017-10-18 05:57 16K

Spark Release 0.3 Spark Release 0.6.0 Spark Release 0.7.0 Spark Release 0.8.0

Spark Release 1.0.0

Spark Release 1.3.0

Spark Release 2.0.0

Spark Release 2.2.0

Apache Spark 2.2.0 is the third release on the 2.x line. This release removes the experimental tag from Structured Streaming. In addition, this release focuses more on usability, stability, and polish, resolving over 1100 tickets.

Additionally, we are excited to announce that PySpark is now available in pypi. To install just run pip install pyspark.

To download Apache Spark 2.2.0, visit the downloads page. You can consult JIRA for the detailed changes. We have curated a list of high level changes here, grouped by major modules.

- Core and Spark SQL
- Structured Streaming
- MLlib
- SparkR

Where does the name 'Spark' come from?



What We Will Cover in This Training

Getting an Environment & Data: CDH + StackOverflow

Refreshing Your Knowledge: Python Fundamentals for This Course

Understanding Spark: An Overview

Getting Technical with Spark



What We Will Cover in This Training

Learning the Core of Spark: RDDs Going Deeper into Spark Core

Increasing Proficiency with Spark: DataFrames & Spark SQL
Continuing the Journey on DataFrames and Spark SQL

Understanding a Typed API: Datasets Works with Scala, Not Python



Picking a Spark Supported Language



Python: easy to use/read, popular, elegant syntax, REPL, many libraries



Scala: type safe, runs on Java VM, concise (good for interactive use)



Java: language of choice for open source and Big Data



R: statistical analysis, many libraries, taught widely, used in Data Science





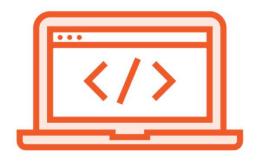
What Do You Need for This Course?



Spark Running Somewhere

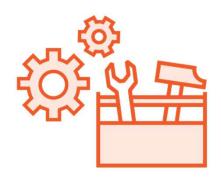








Why Cloudera?







DirectorCloud



AltusPlatform-as-a-Service



Creating Your First Big Data Hadoop Cluster Using Cloudera CDH

★★★★ By Xavier Morer

Data by itself has no meaning, it is what you do with it that counts. In this course, you'll fast track to Hadoop & Big Data with the Cloudera QuickStart VM and then you'll learn how to set up a Hadoop cluster with Cloudera CDH.

Preparing a Production Hadoop Cluster with Cloudera: Databases

By Xavier Morera

Big Data is a natural evolution of data analysis, scaling beyond the limits of conventional databases. However, they're still an important part of a Hadoop cluster. Learn how to setup databases for Cloudera CDH and install a production grade cluster.

A few courses that might help you

Deploying Hadoop with Cloudera CDH to AWS

By Xavier Morera

Learn how to deploy, size, and scale Hadoop in the cloud (namely AWS). You'll understand key concepts to deploy a CDH cluster, perform a manual installation, and finally learn how to automate deployments for multiple clusters with Cloudera Director.

Take Control of Your Big Data with HUE in Cloudera CDH

By Xavier Morera

Working with Big Data is no small task. Jumpstart your Hadoop skills by loading, visualizing, analyzing, and searching your data using Cloudera HUE, the Hadoop User Experience. Take control of your Big Data!

Takeaway



Spark + Python + Cloudera

History of Spark: From RDDs to today

Overview of what's coming

A few related courses that might help