Getting an Environment & Data: CDH + StackOverflow

Xavier Morera

HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

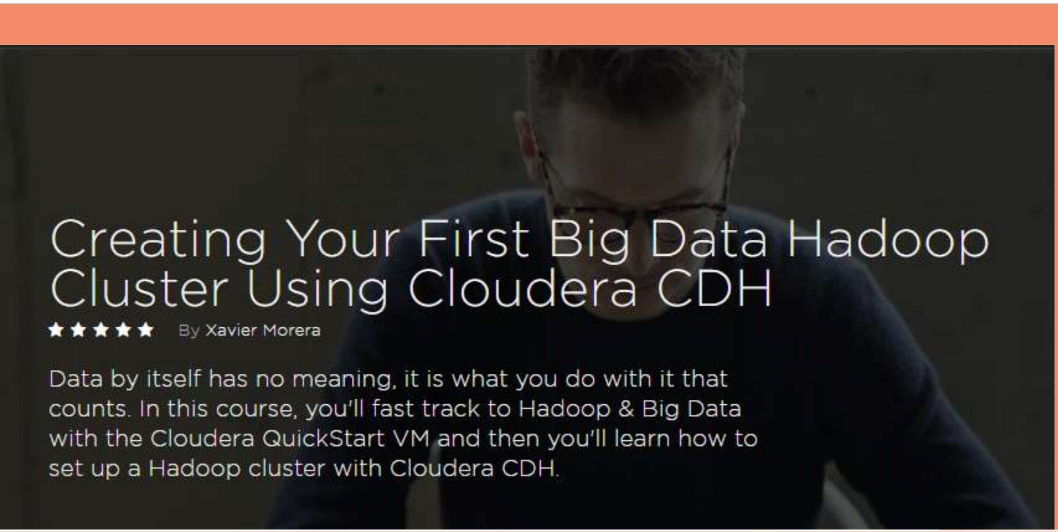
@xmorera www.xaviermorera.com

Getting an Environment & Data

CDH

Cloudera's Distribution including Hadoop

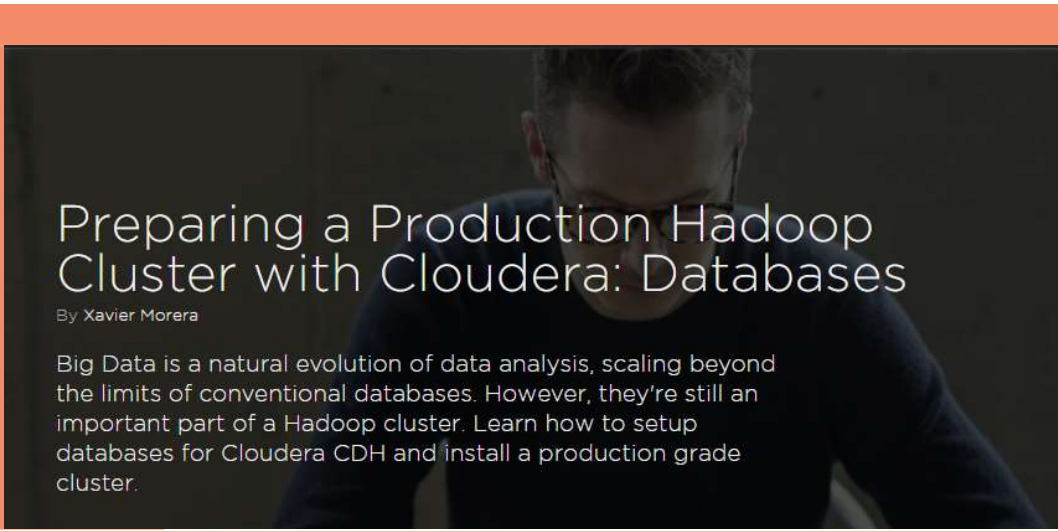
Getting a Cloudera Cluster



Creating Your First Big Data Hadoop Cluster Using Cloudera CDH

★★★★★ By Xavier Morera

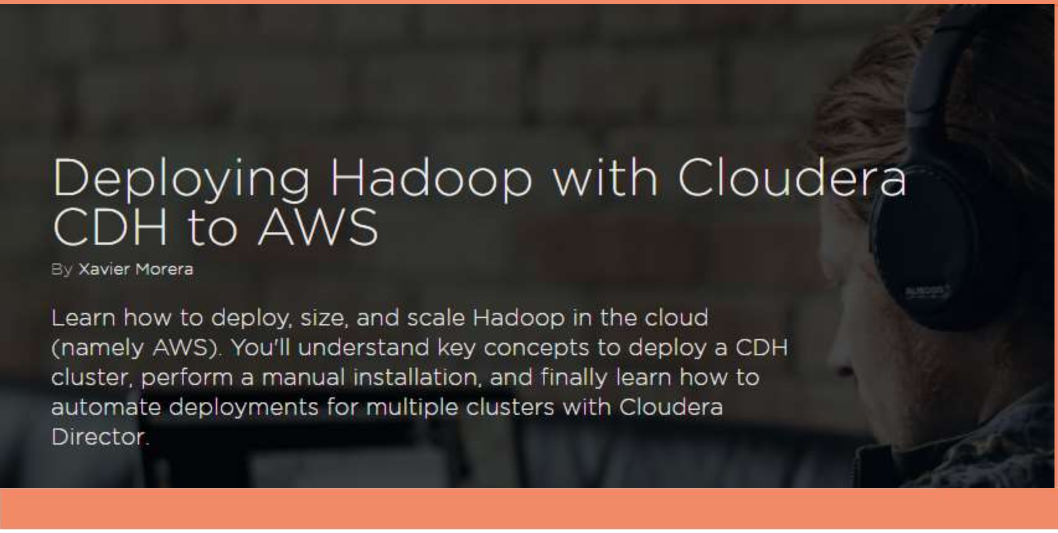
Data by itself has no meaning, it is what you do with it that counts. In this course, you'll fast track to Hadoop & Big Data with the Cloudera QuickStart VM and then you'll learn how to set up a Hadoop cluster with Cloudera CDH.



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.



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.

Getting an Environment & Data

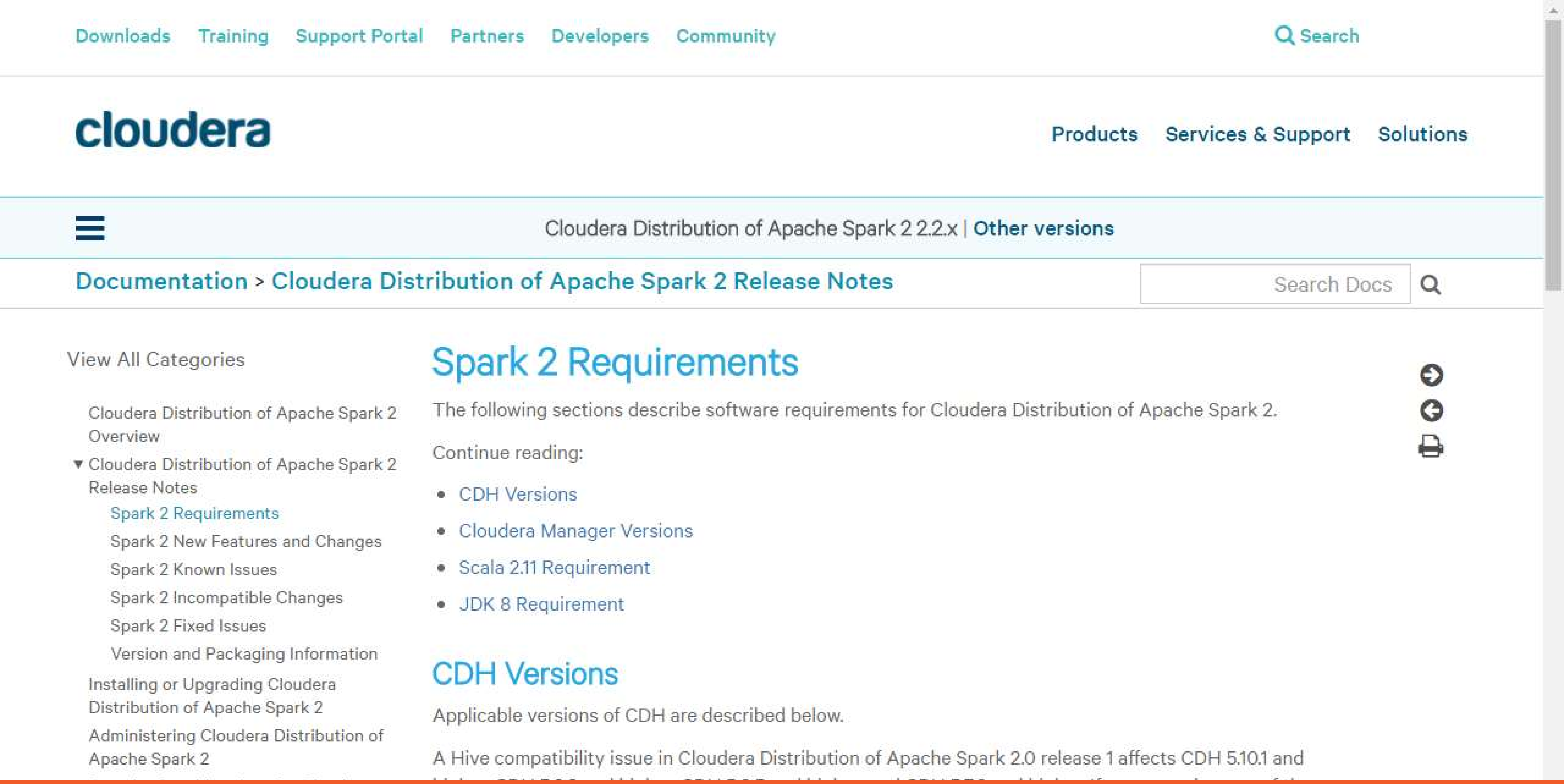
CDH

Cloudera's Distribution including Hadoop

StackOverflow

Question & Answer site for programmers

Prerequisites & Known Issues


Downloads
Training
Support Portal
Partners
Developers
Community
Search
cloudera
Products
Services & Support
Solutions
Cloudera Distribution of Apache Spark 2 2.2.x | Other versions
Documentation > Cloudera Distribution of Apache Spark 2 Release Notes
Search Docs
View All Categories
Cloudera Distribution of Apache Spark 2 Overview
▼ Cloudera Distribution of Apache Spark 2 Release Notes
Spark 2 Requirements
Spark 2 New Features and Changes
Spark 2 Known Issues
Spark 2 Incompatible Changes
Spark 2 Fixed Issues
Version and Packaging Information
Installing or Upgrading Cloudera Distribution of Apache Spark 2
Administering Cloudera Distribution of Apache Spark 2

Spark 2 Requirements

The following sections describe software requirements for Cloudera Distribution of Apache Spark 2.

Continue reading:

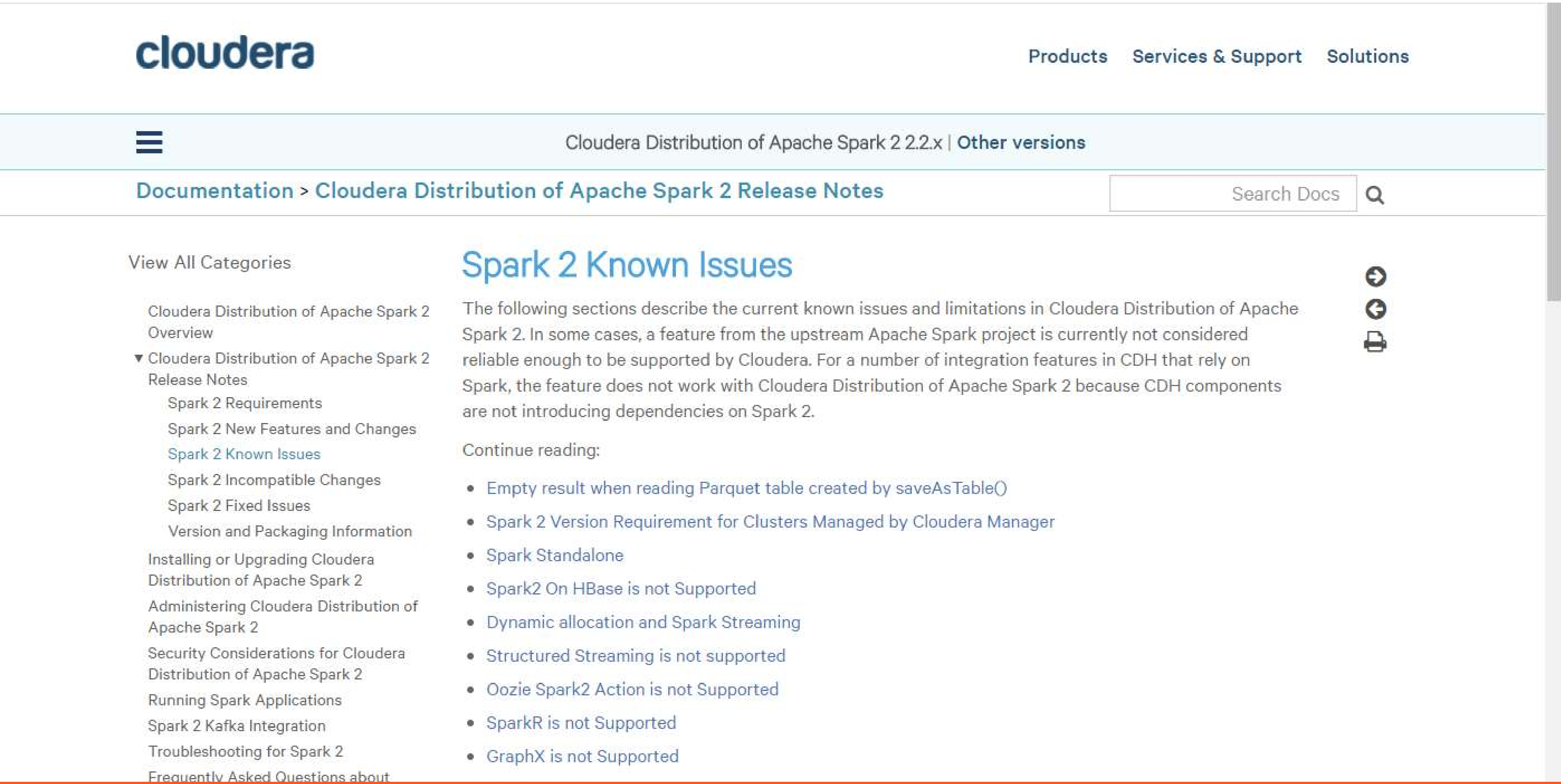
- CDH Versions
- Cloudera Manager Versions
- Scala 2.11 Requirement
- JDK 8 Requirement

CDH Versions

Applicable versions of CDH are described below.

A Hive compatibility issue in Cloudera Distribution of Apache Spark 2.0 release 1 affects CDH 5.10.1 and

http://tiny.bigdatainc.org/sparkreq


cloudera
Products
Services & Support
Solutions
Cloudera Distribution of Apache Spark 2 2.2.x | Other versions
Documentation > Cloudera Distribution of Apache Spark 2 Release Notes
Search Docs
View All Categories
Cloudera Distribution of Apache Spark 2 Overview
▼ Cloudera Distribution of Apache Spark 2 Release Notes
Spark 2 Requirements
Spark 2 New Features and Changes
Spark 2 Known Issues
Spark 2 Incompatible Changes
Spark 2 Fixed Issues
Version and Packaging Information
Installing or Upgrading Cloudera Distribution of Apache Spark 2
Administering Cloudera Distribution of Apache Spark 2
Security Considerations for Cloudera Distribution of Apache Spark 2
Running Spark Applications
Spark 2 Kafka Integration
Troubleshooting for Spark 2
Frequently Asked Questions about

Spark 2 Known Issues

The following sections describe the current known issues and limitations in Cloudera Distribution of Apache Spark 2. In some cases, a feature from the upstream Apache Spark project is currently not considered reliable enough to be supported by Cloudera. For a number of integration features in CDH that rely on Spark, the feature does not work with Cloudera Distribution of Apache Spark 2 because CDH components are not introducing dependencies on Spark 2.

Continue reading:

- Empty result when reading Parquet table created by saveAsTable()
- Spark 2 Version Requirement for Clusters Managed by Cloudera Manager
- Spark Standalone
- Spark2 On HBase is not Supported
- Dynamic allocation and Spark Streaming
- Structured Streaming is not supported
- Oozie Spark2 Action is not Supported
- SparkR is not Supported
- GraphX is not Supported

http://tiny.bigdatainc.org/spark2ki

What's Needed for Spark 2

5.8.3+	5.8+	2.11	8
CDH	Cloudera Manager	Scala	JDK

Spark 1.6 can coexist with 2.x

Upgrading Cloudera Manager & CDH

Upgrading Cloudera Manager & CDH

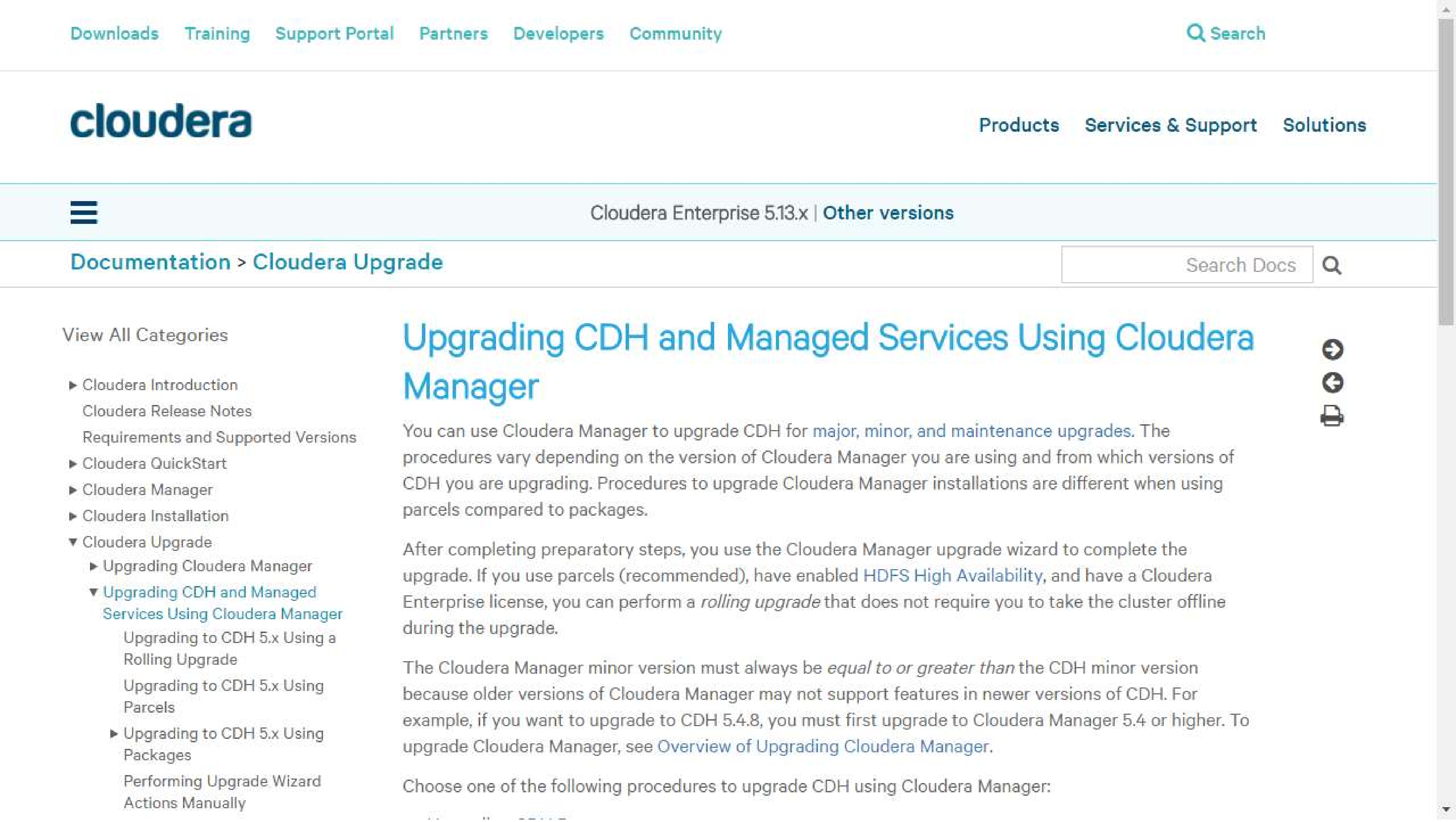
5.8+

CDH

5.8.3, 5.9+

Cloudera Manager

http://tiny.bigdatainc.org/upcdh



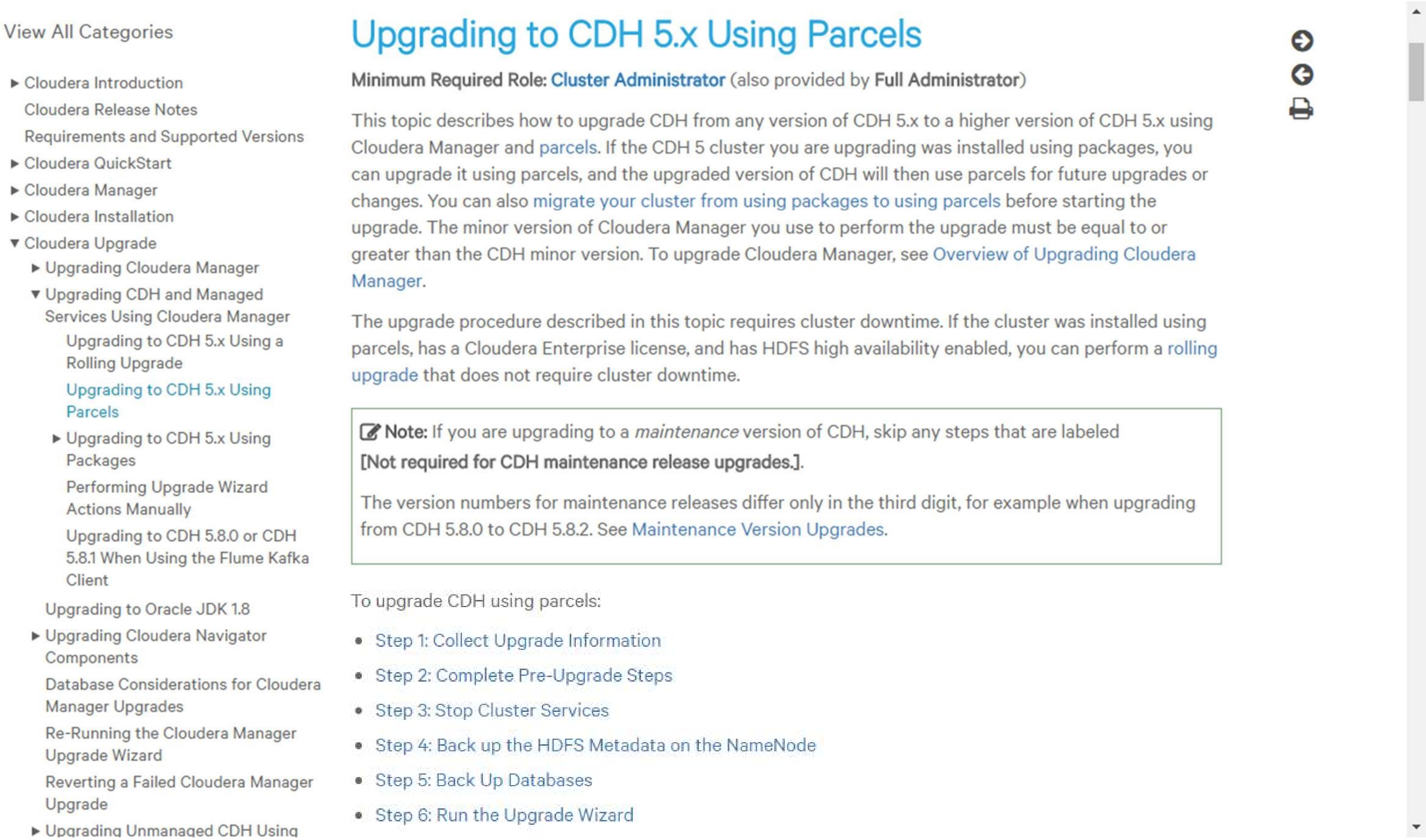
Upgrading CDH and Managed Services Using Cloudera Manager

You can use Cloudera Manager to upgrade CDH for major, minor, and maintenance upgrades. The procedures vary depending on the version of Cloudera Manager you are using and from which versions of CDH you are upgrading. Procedures to upgrade Cloudera Manager installations are different when using parcels compared to packages.

After completing preparatory steps, you use the Cloudera Manager upgrade wizard to complete the upgrade. If you use parcels (recommended), have enabled HDFS High Availability, and have a Cloudera Enterprise license, you can perform a *rolling upgrade* that does not require you to take the cluster offline during the upgrade.

The Cloudera Manager minor version must always be *equal to or greater than* the CDH minor version because older versions of Cloudera Manager may not support features in newer versions of CDH. For example, if you want to upgrade to CDH 5.4.8, you must first upgrade to Cloudera Manager 5.4 or higher. To upgrade Cloudera Manager, see Overview of Upgrading Cloudera Manager.

Choose one of the following procedures to upgrade CDH using Cloudera Manager:



View All Categories

- ▶ Cloudera Introduction
- Cloudera Release Notes
- Requirements and Supported Versions
- ▶ Cloudera QuickStart
- ▶ Cloudera Manager
- ▶ Cloudera Installation
- ▼ Cloudera Upgrade
  - ▶ Upgrading Cloudera Manager
  - ▼ Upgrading CDH and Managed Services Using Cloudera Manager
    - Upgrading to CDH 5.x Using a Rolling Upgrade
    - Upgrading to CDH 5.x Using Parcels
    - ▶ Upgrading to CDH 5.x Using Packages
    - Performing Upgrade Wizard Actions Manually
    - Upgrading to CDH 5.8.0 or CDH 5.8.1 When Using the Flume Kafka Client
  - Upgrading to Oracle JDK 1.8
  - ▶ Upgrading Cloudera Navigator Components
  - Database Considerations for Cloudera Manager Upgrades
  - Re-Running the Cloudera Manager Upgrade Wizard
  - Reverting a Failed Cloudera Manager Upgrade
  - ▶ Upgrading Unmanaged CDH Using

Upgrading to CDH 5.x Using Parcels

Minimum Required Role: Cluster Administrator (also provided by **Full Administrator**)

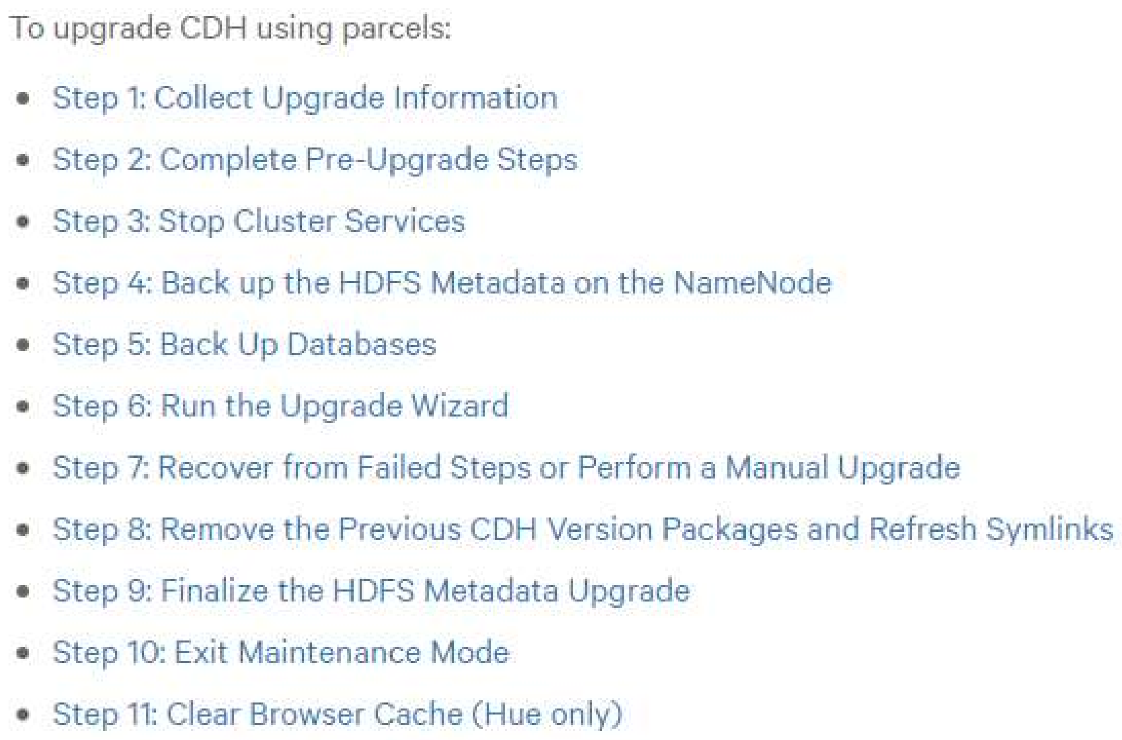
This topic describes how to upgrade CDH from any version of CDH 5.x to a higher version of CDH 5.x using Cloudera Manager and parcels. If the CDH 5 cluster you are upgrading was installed using packages, you can upgrade it using parcels, and the upgraded version of CDH will then use parcels for future upgrades or changes. You can also migrate your cluster from using packages to using parcels before starting the upgrade. The minor version of Cloudera Manager you use to perform the upgrade must be equal to or greater than the CDH minor version. To upgrade Cloudera Manager, see Overview of Upgrading Cloudera Manager.

The upgrade procedure described in this topic requires cluster downtime. If the cluster was installed using parcels, has a Cloudera Enterprise license, and has HDFS high availability enabled, you can perform a rolling upgrade that does not require cluster downtime.

> **Note:** If you are upgrading to a *maintenance* version of CDH, skip any steps that are labeled **[Not required for CDH maintenance release upgrades.]**.
>
> The version numbers for maintenance releases differ only in the third digit, for example when upgrading from CDH 5.8.0 to CDH 5.8.2. See Maintenance Version Upgrades.

To upgrade CDH using parcels:

- Step 1: Collect Upgrade Information
- Step 2: Complete Pre-Upgrade Steps
- Step 3: Stop Cluster Services
- Step 4: Back up the HDFS Metadata on the NameNode
- Step 5: Back Up Databases
- Step 6: Run the Upgrade Wizard



To upgrade CDH using parcels:

- Step 1: Collect Upgrade Information
- Step 2: Complete Pre-Upgrade Steps
- Step 3: Stop Cluster Services
- Step 4: Back up the HDFS Metadata on the NameNode
- Step 5: Back Up Databases
- Step 6: Run the Upgrade Wizard
- Step 7: Recover from Failed Steps or Perform a Manual Upgrade
- Step 8: Remove the Previous CDH Version Packages and Refresh Symlinks
- Step 9: Finalize the HDFS Metadata Upgrade
- Step 10: Exit Maintenance Mode
- Step 11: Clear Browser Cache (Hue only)

Installing or Upgrading JDK 1.8

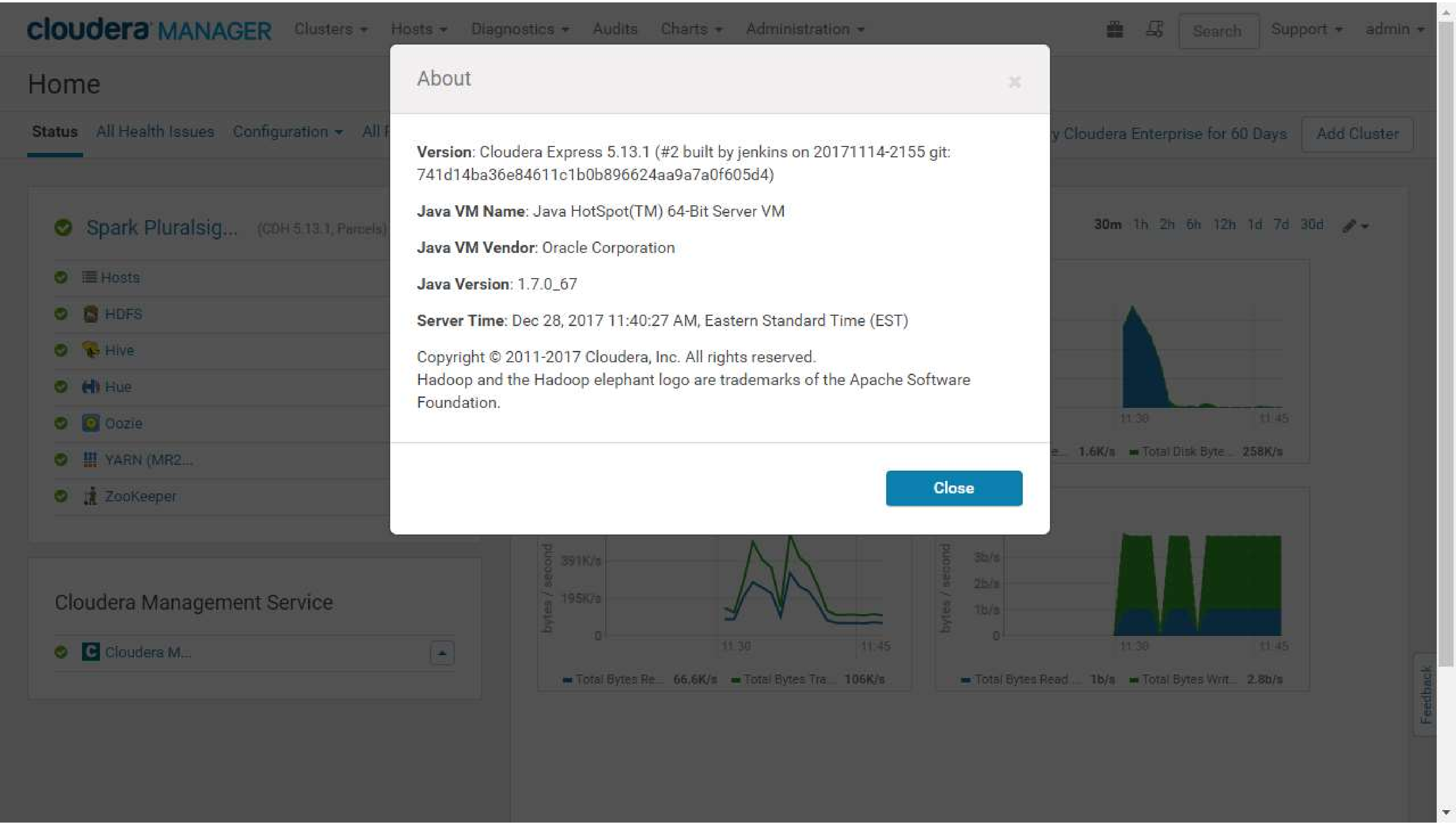
Installing or Upgrading JDK 1.8

Possible that your cluster has JDK 1.7

Spark 2.x requires JDK 1.8

JDK 8 supported since CDH 5.3

- A few ways to get JDK 1.8
- Spark 2.x requires specific versions



Version: Cloudera Express 5.13.1 (#2 built by jenkins on 20171114-2155 git: 741d14ba36e84611c1b0b896624aa9a7a0f605d4)

Java VM Name: Java HotSpot(TM) 64-Bit Server VM

Java VM Vendor: Oracle Corporation

Java Version: 1.7.0_67

Server Time: Dec 28, 2017 11:40:27 AM, Eastern Standard Time (EST)

Copyright © 2011-2017 Cloudera, Inc. All rights reserved.
Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation.

Getting Spark with CDH

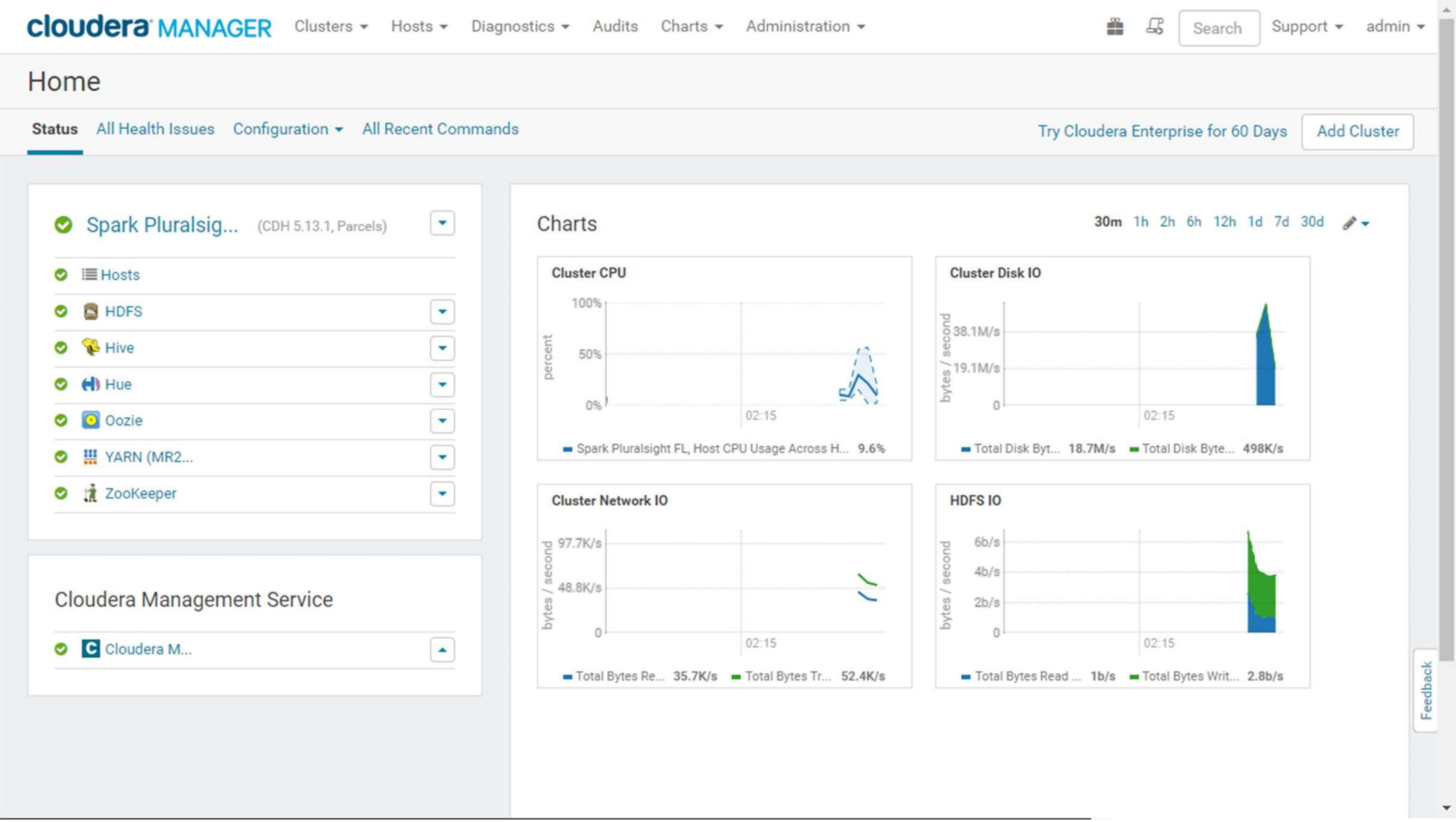
Getting Spark with CDH

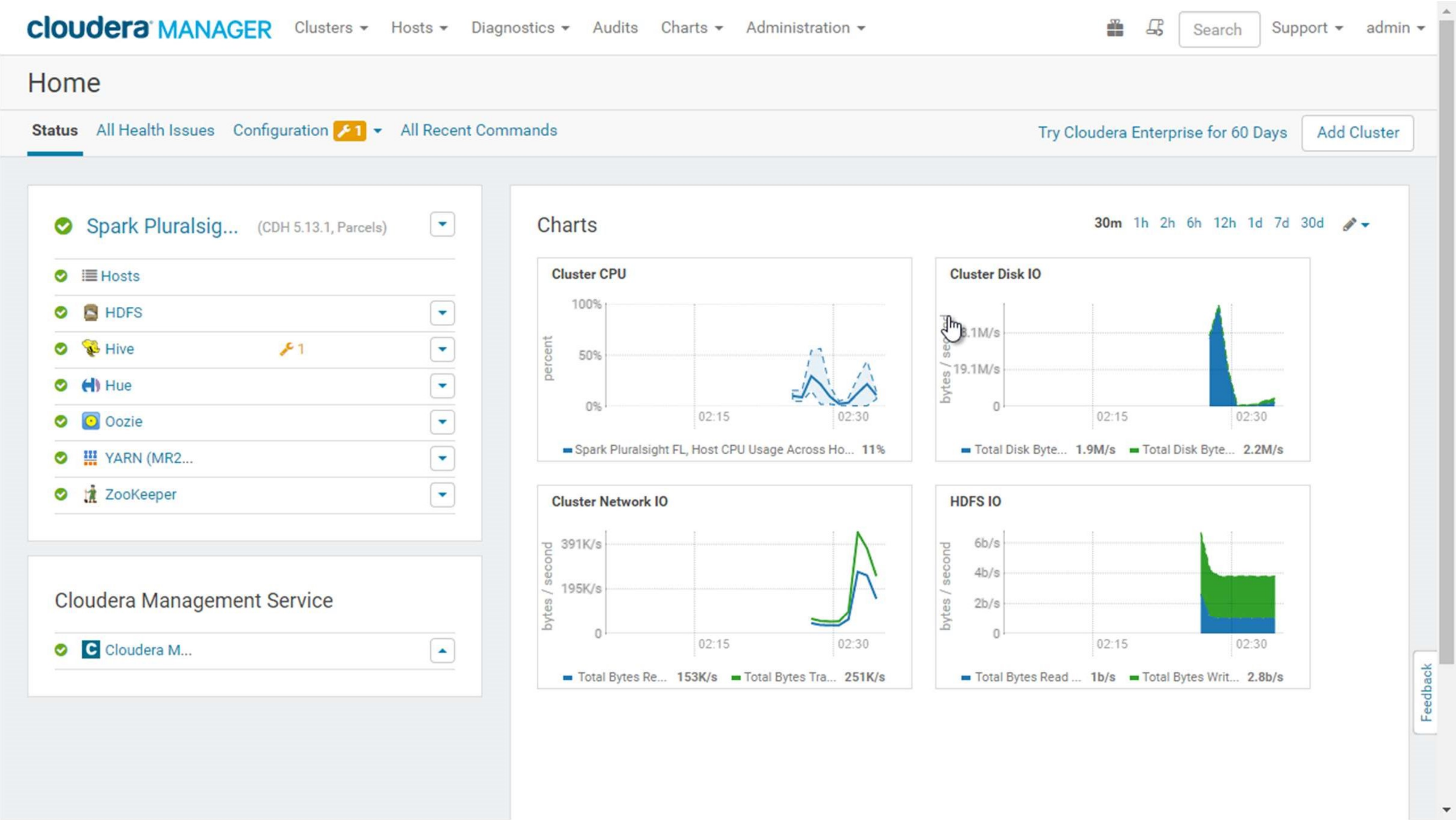
Spark 1.6

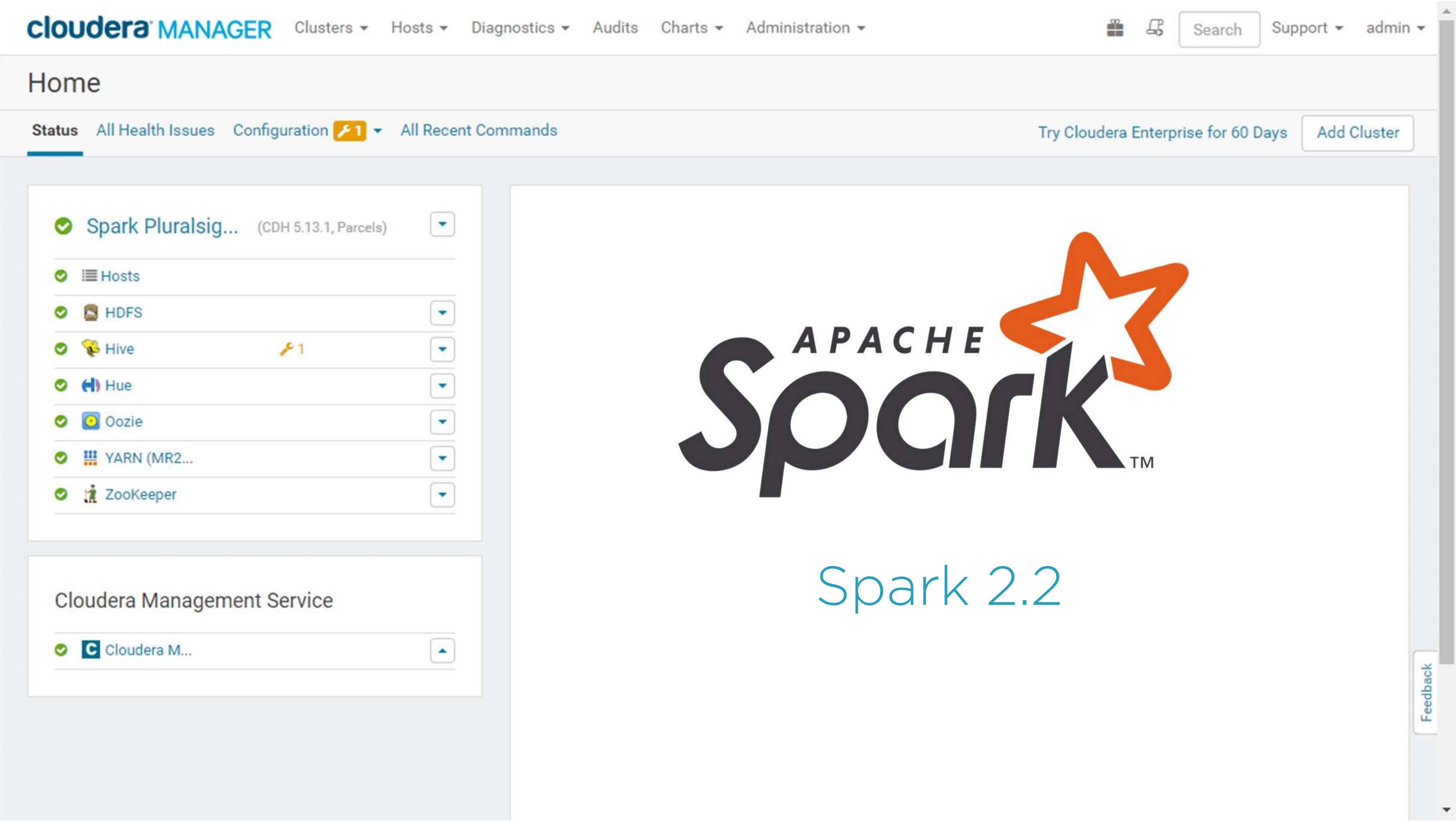
Spark 2.2

Getting Spark 1.6 with CDH

Spark 1.6


cloudera MANAGER
Clusters
Hosts
Diagnostics
Audits
Charts
Administration
Search
Support
admin
Home
Status
All Health Issues
Configuration
All Recent Commands
Try Cloudera Enterprise for 60 Days
Add Cluster
Spark Pluralsig... (CDH 5.13.1, Parcels)
Hosts
HDFS
Hive
Hue
Oozie
YARN (MR2...
ZooKeeper
Cloudera Management Service
Cloudera M...
Charts
30m
1h
2h
6h
12h
1d
7d
30d
Cluster CPU
100%
50%
0%
percent
02:15
Spark Pluralsight FL, Host CPU Usage Across H... 9.6%
Cluster Disk IO
38.1M/s
19.1M/s
0
bytes / second
02:15
Total Disk Byt... 18.7M/s
Total Disk Byte... 498K/s
Cluster Network IO
97.7K/s
48.8K/s
0
bytes / second
02:15
Total Bytes Re... 35.7K/s
Total Bytes Tr... 52.4K/s
HDFS IO
6b/s
4b/s
2b/s
0
bytes / second
02:15
Total Bytes Read ... 1b/s
Total Bytes Writ... 2.8b/s
Feedback


cloudera MANAGER
Clusters
Hosts
Diagnostics
Audits
Charts
Administration
Search
Support
admin
Home
Status
All Health Issues
Configuration 1
All Recent Commands
Try Cloudera Enterprise for 60 Days
Add Cluster
Spark Pluralsig... (CDH 5.13.1, Parcels)
Hosts
HDFS
Hive 1
Hue
Oozie
YARN (MR2...
ZooKeeper
Cloudera Management Service
Cloudera M...
Charts
30m 1h 2h 6h 12h 1d 7d 30d
Cluster CPU
100%
50%
0%
percent
02:15
02:30
Spark Pluralsight FL, Host CPU Usage Across Ho... 11%
Cluster Disk IO
38.1M/s
19.1M/s
0
bytes / second
02:15
02:30
Total Disk Byte... 1.9M/s
Total Disk Byte... 2.2M/s
Cluster Network IO
391K/s
195K/s
0
bytes / second
02:15
02:30
Total Bytes Re... 153K/s
Total Bytes Tra... 251K/s
HDFS IO
6b/s
4b/s
2b/s
0
bytes / second
02:15
02:30
Total Bytes Read ... 1b/s
Total Bytes Writ... 2.8b/s
Feedback


cloudera MANAGER
Clusters
Hosts
Diagnostics
Audits
Charts
Administration
Search
Support
admin
Home
Status
All Health Issues
Configuration 1
All Recent Commands
Try Cloudera Enterprise for 60 Days
Add Cluster
Spark Pluralsig... (CDH 5.13.1, Parcels)
Hosts
HDFS
Hive 1
Hue
Oozie
YARN (MR2...
ZooKeeper
Cloudera Management Service
Cloudera M...
APACHE Spark ™
Feedback

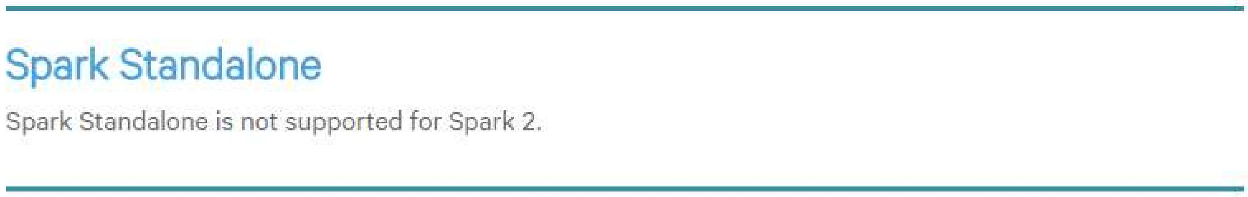
Spark 2.2

Getting Spark 2 Standalone

Getting Spark 2 Standalone

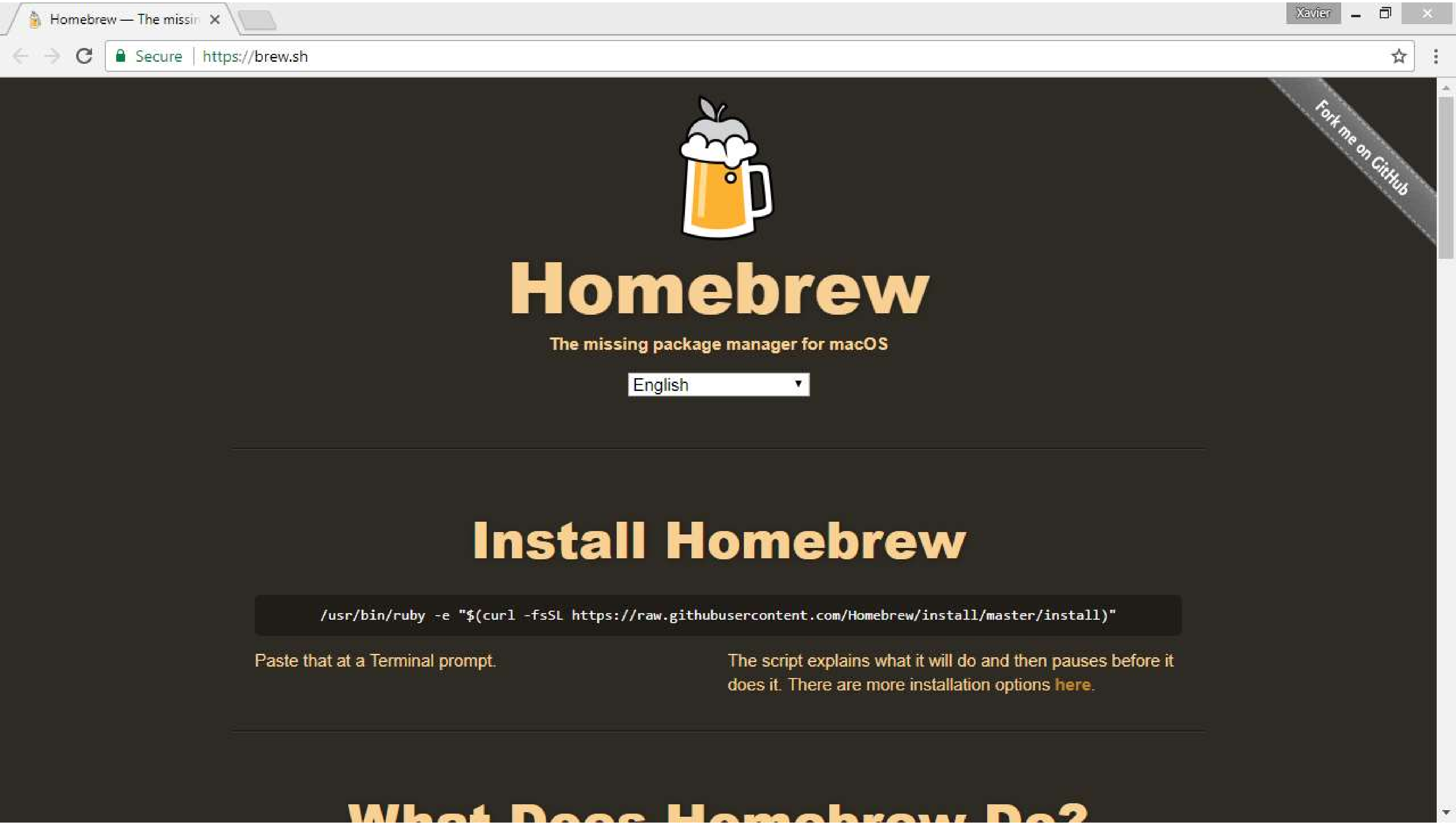
Standalone

Spark 2.2



Spark Standalone

Spark Standalone is not supported for Spark 2.



Homebrew

The missing package manager for macOS

Install Homebrew

```
/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"
```

Paste that at a Terminal prompt.

The script explains what it will do and then pauses before it does it. There are more installation options here.

What Does Homebrew Do?

Installing Spark 2 on Cloudera

Installing Spark 2 on Cloudera

Prerequisites

Download Spark2 CSD and install

- Custom Service Descriptor

Add Spark2 Parcel Repository

Download, Distribute & Activate

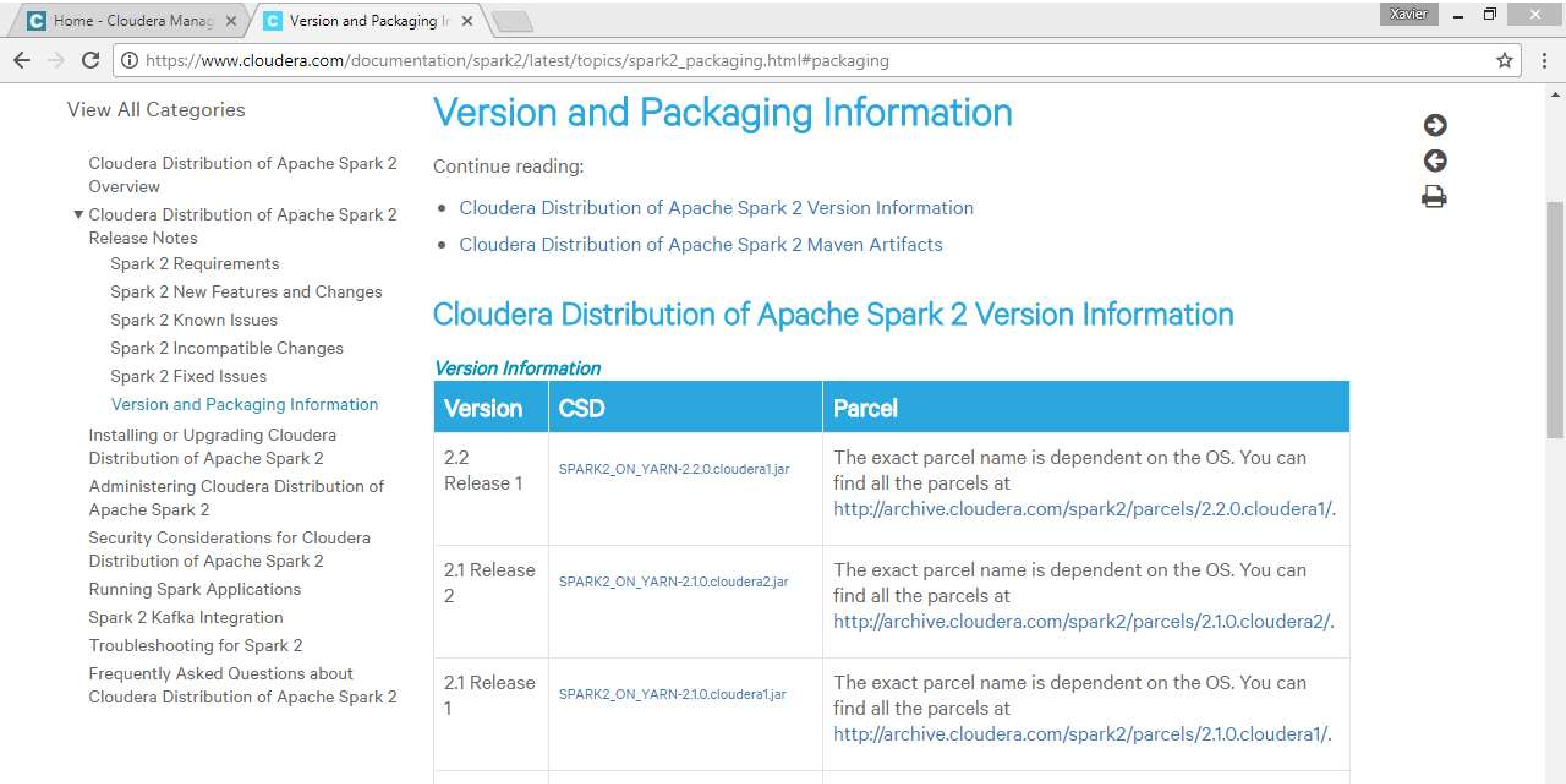
Installing Spark 2 on Cloudera

Add Spark 2 Service

Restart Stale Services

Redeploy Client Configuration

Spark 2 away!


Home - Cloudera Manag
Version and Packaging Ir
Xavier
https://www.cloudera.com/documentation/spark2/latest/topics/spark2_packaging.html#packaging

View All Categories

- Cloudera Distribution of Apache Spark 2 Overview
- ▼ Cloudera Distribution of Apache Spark 2 Release Notes
  - Spark 2 Requirements
  - Spark 2 New Features and Changes
  - Spark 2 Known Issues
  - Spark 2 Incompatible Changes
  - Spark 2 Fixed Issues
  - Version and Packaging Information
- Installing or Upgrading Cloudera Distribution of Apache Spark 2
- Administering Cloudera Distribution of Apache Spark 2
- Security Considerations for Cloudera Distribution of Apache Spark 2
- Running Spark Applications
- Spark 2 Kafka Integration
- Troubleshooting for Spark 2
- Frequently Asked Questions about Cloudera Distribution of Apache Spark 2

Version and Packaging Information

Continue reading:

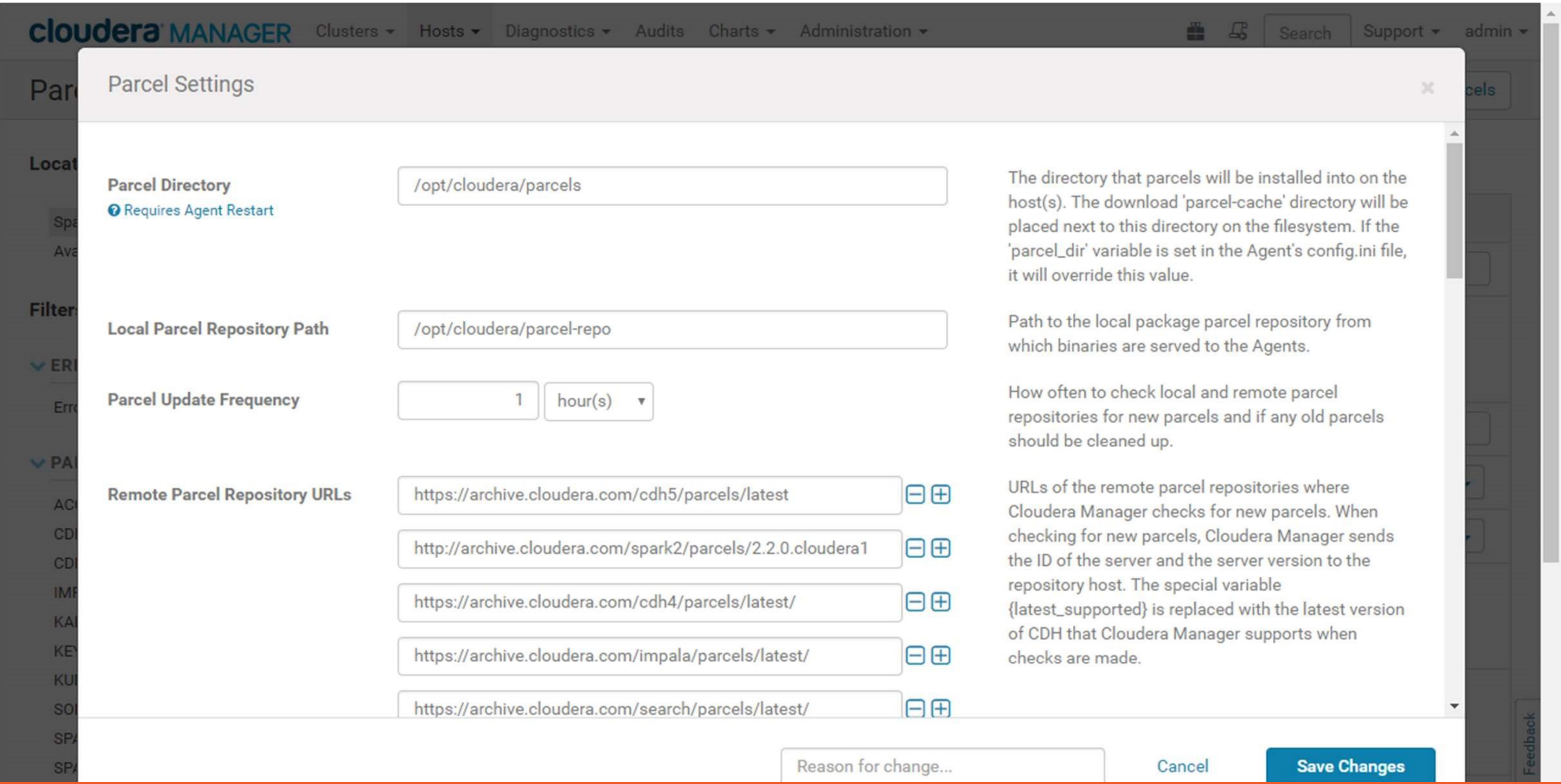
- Cloudera Distribution of Apache Spark 2 Version Information
- Cloudera Distribution of Apache Spark 2 Maven Artifacts

Cloudera Distribution of Apache Spark 2 Version Information

Version Information

Version	CSD	Parcel
2.2 Release 1	SPARK2_ON_YARN-2.2.0.cloudera1.jar	The exact parcel name is dependent on the OS. You can find all the parcels at http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera1/.
2.1 Release 2	SPARK2_ON_YARN-2.1.0.cloudera2.jar	The exact parcel name is dependent on the OS. You can find all the parcels at http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera2/.
2.1 Release 1	SPARK2_ON_YARN-2.1.0.cloudera1.jar	The exact parcel name is dependent on the OS. You can find all the parcels at http://archive.cloudera.com/spark2/parcels/2.1.0.cloudera1/.

http://tiny.bigdatainc.org/spark2csd



Parcel Settings

Parcel Directory
Requires Agent Restart

/opt/cloudera/parcels

The directory that parcels will be installed into on the host(s). The download 'parcel-cache' directory will be placed next to this directory on the filesystem. If the 'parcel_dir' variable is set in the Agent's config.ini file, it will override this value.

Local Parcel Repository Path

/opt/cloudera/parcel-repo

Path to the local package parcel repository from which binaries are served to the Agents.

Parcel Update Frequency

1 hour(s)

How often to check local and remote parcel repositories for new parcels and if any old parcels should be cleaned up.

Remote Parcel Repository URLs

https://archive.cloudera.com/cdh5/parcels/latest
http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera1
https://archive.cloudera.com/cdh4/parcels/latest/
https://archive.cloudera.com/impala/parcels/latest/
https://archive.cloudera.com/search/parcels/latest/

URLs of the remote parcel repositories where Cloudera Manager checks for new parcels. When checking for new parcels, Cloudera Manager sends the ID of the server and the server version to the repository host. The special variable {latest_supported} is replaced with the latest version of CDH that Cloudera Manager supports when checks are made.

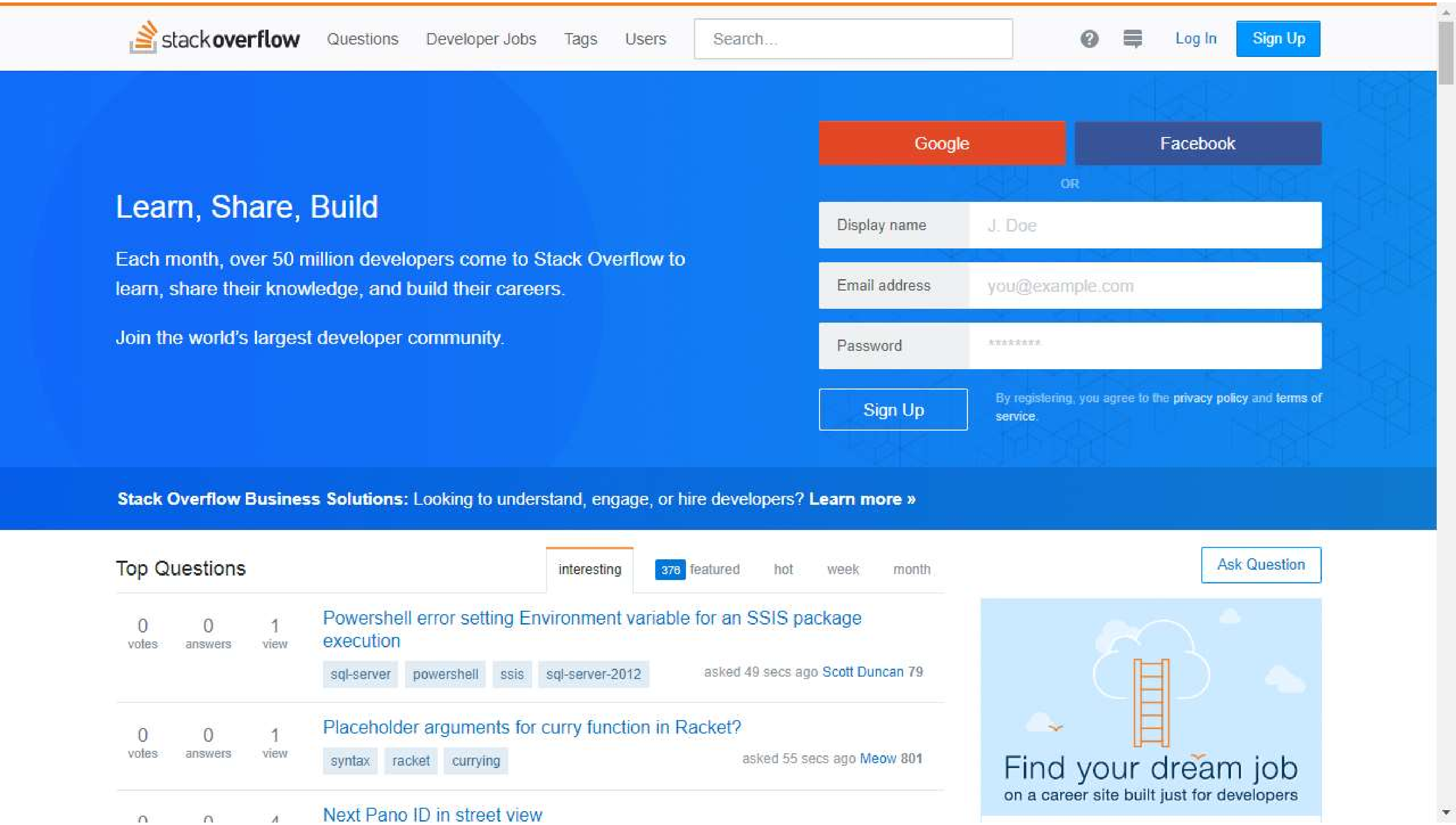
Reason for change...

Cancel

Save Changes

http://tiny.bigdatainc.org/spark2parcel

Your Big Data
StackOverflow & StackExchange



Learn, Share, Build

Each month, over 50 million developers come to Stack Overflow to learn, share their knowledge, and build their careers.

Join the world's largest developer community.

Display name: J. Doe

Email address: you@example.com

Password: ********

By registering, you agree to the **privacy policy** and **terms of service**.

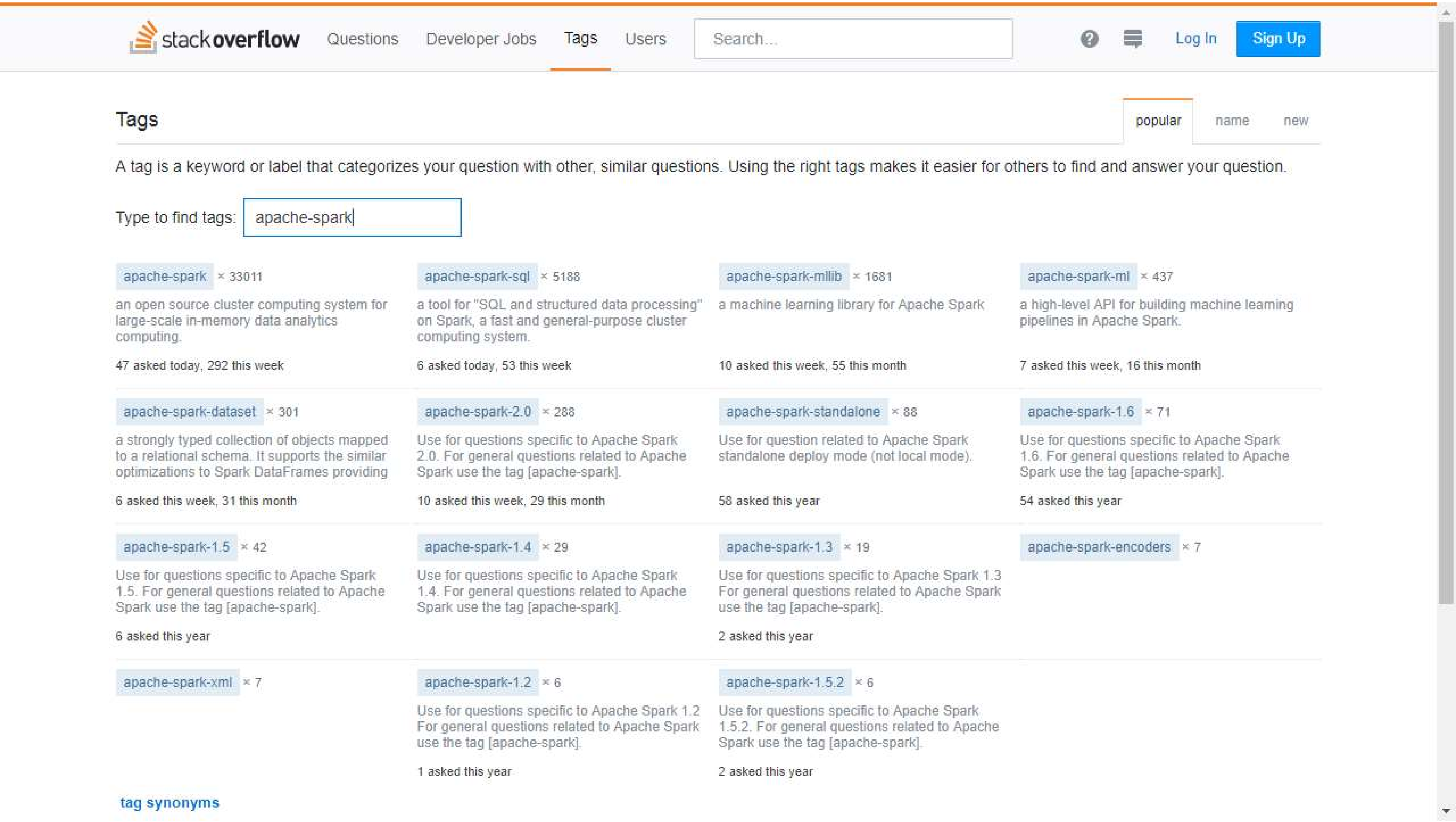
Stack Overflow Business Solutions: Looking to understand, engage, or hire developers? **Learn more »**

Top Questions

0 votes 0 answers 1 view — Powershell error setting Environment variable for an SSIS package execution
sql-server powershell ssis sql-server-2012
asked 49 secs ago Scott Duncan 79

0 votes 0 answers 1 view — Placeholder arguments for curry function in Racket?
syntax racket currying
asked 55 secs ago Meow 801

0 0 1 Next Pano ID in street view



Tags

A tag is a keyword or label that categorizes your question with other, similar questions. Using the right tags makes it easier for others to find and answer your question.

Type to find tags: apache-spark

apache-spark × 33011
an open source cluster computing system for large-scale in-memory data analytics computing.
47 asked today, 292 this week

apache-spark-sql × 5188
a tool for "SQL and structured data processing" on Spark, a fast and general-purpose cluster computing system.
6 asked today, 53 this week

apache-spark-mllib × 1681
a machine learning library for Apache Spark
10 asked this week, 55 this month

apache-spark-ml × 437
a high-level API for building machine learning pipelines in Apache Spark.
7 asked this week, 16 this month

apache-spark-dataset × 301
a strongly typed collection of objects mapped to a relational schema. It supports the similar optimizations to Spark DataFrames providing
6 asked this week, 31 this month

apache-spark-2.0 × 288
Use for questions specific to Apache Spark 2.0. For general questions related to Apache Spark use the tag [apache-spark].
10 asked this week, 29 this month

apache-spark-standalone × 88
Use for question related to Apache Spark standalone deploy mode (not local mode).
58 asked this year

apache-spark-1.6 × 71
Use for questions specific to Apache Spark 1.6. For general questions related to Apache Spark use the tag [apache-spark].
54 asked this year

apache-spark-1.5 × 42
Use for questions specific to Apache Spark 1.5. For general questions related to Apache Spark use the tag [apache-spark].
6 asked this year

apache-spark-1.4 × 29
Use for questions specific to Apache Spark 1.4. For general questions related to Apache Spark use the tag [apache-spark].

apache-spark-1.3 × 19
Use for questions specific to Apache Spark 1.3 For general questions related to Apache Spark use the tag [apache-spark].
2 asked this year

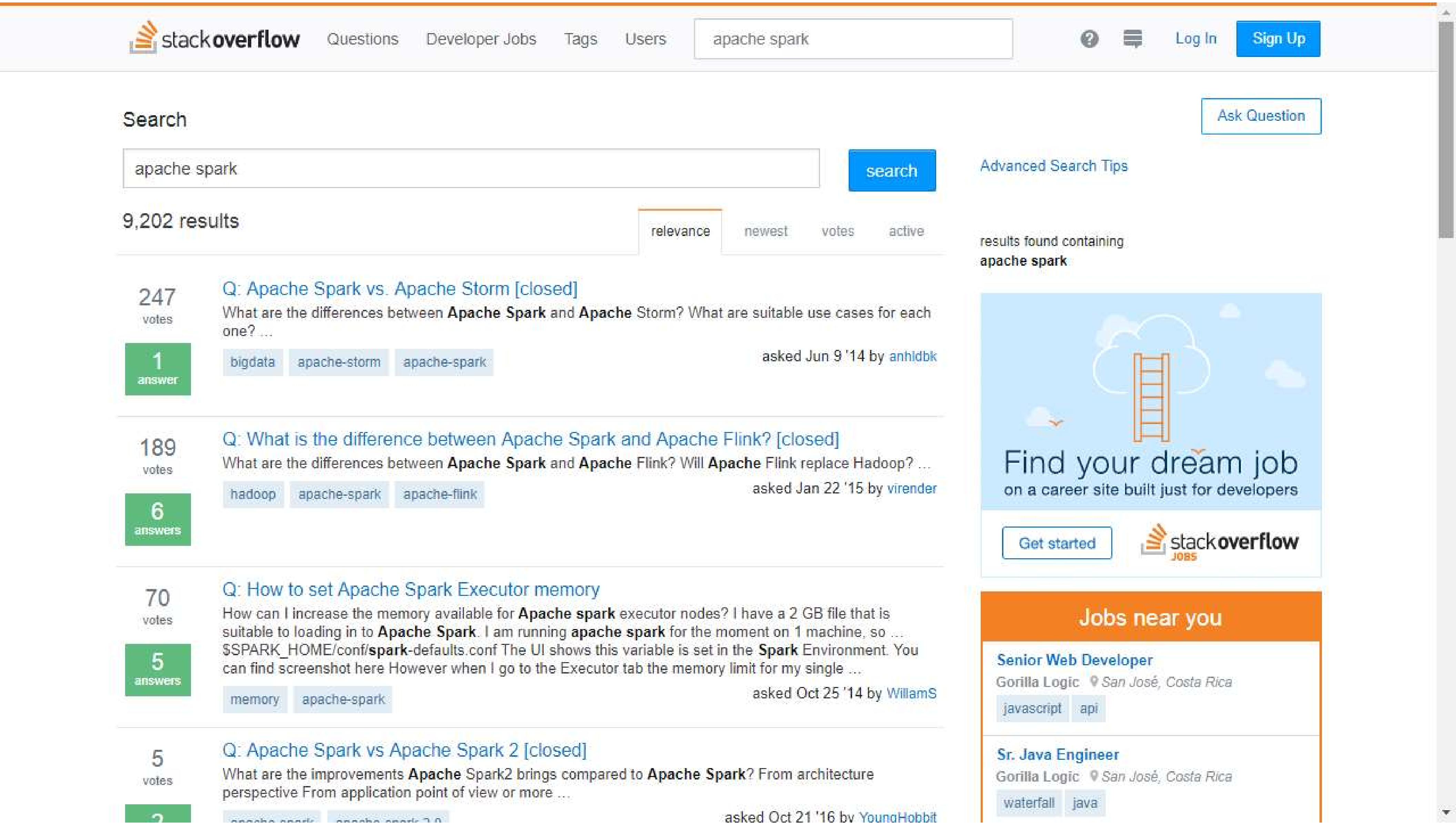
apache-spark-encoders × 7

apache-spark-xml × 7

apache-spark-1.2 × 6
Use for questions specific to Apache Spark 1.2 For general questions related to Apache Spark use the tag [apache-spark].
1 asked this year

apache-spark-1.5.2 × 6
Use for questions specific to Apache Spark 1.5.2. For general questions related to Apache Spark use the tag [apache-spark].
2 asked this year

tag synonyms



Search

apache spark

Advanced Search Tips

9,202 results

results found containing **apache spark**

247 votes
1 answer

Q: Apache Spark vs. Apache Storm [closed]

What are the differences between **Apache Spark** and **Apache** Storm? What are suitable use cases for each one? ...

bigdata apache-storm apache-spark

asked Jun 9 '14 by anhldbk

189 votes
6 answers

Q: What is the difference between Apache Spark and Apache Flink? [closed]

What are the differences between **Apache Spark** and **Apache** Flink? Will **Apache** Flink replace Hadoop? ...

hadoop apache-spark apache-flink

asked Jan 22 '15 by virender

70 votes
5 answers

Q: How to set Apache Spark Executor memory

How can I increase the memory available for **Apache spark** executor nodes? I have a 2 GB file that is suitable to loading in to **Apache Spark**. I am running **apache spark** for the moment on 1 machine, so ... $SPARK_HOME/conf/**spark**-defaults.conf The UI shows this variable is set in the **Spark** Environment. You can find screenshot here However when I go to the Executor tab the memory limit for my single ...

memory apache-spark

asked Oct 25 '14 by WillamS

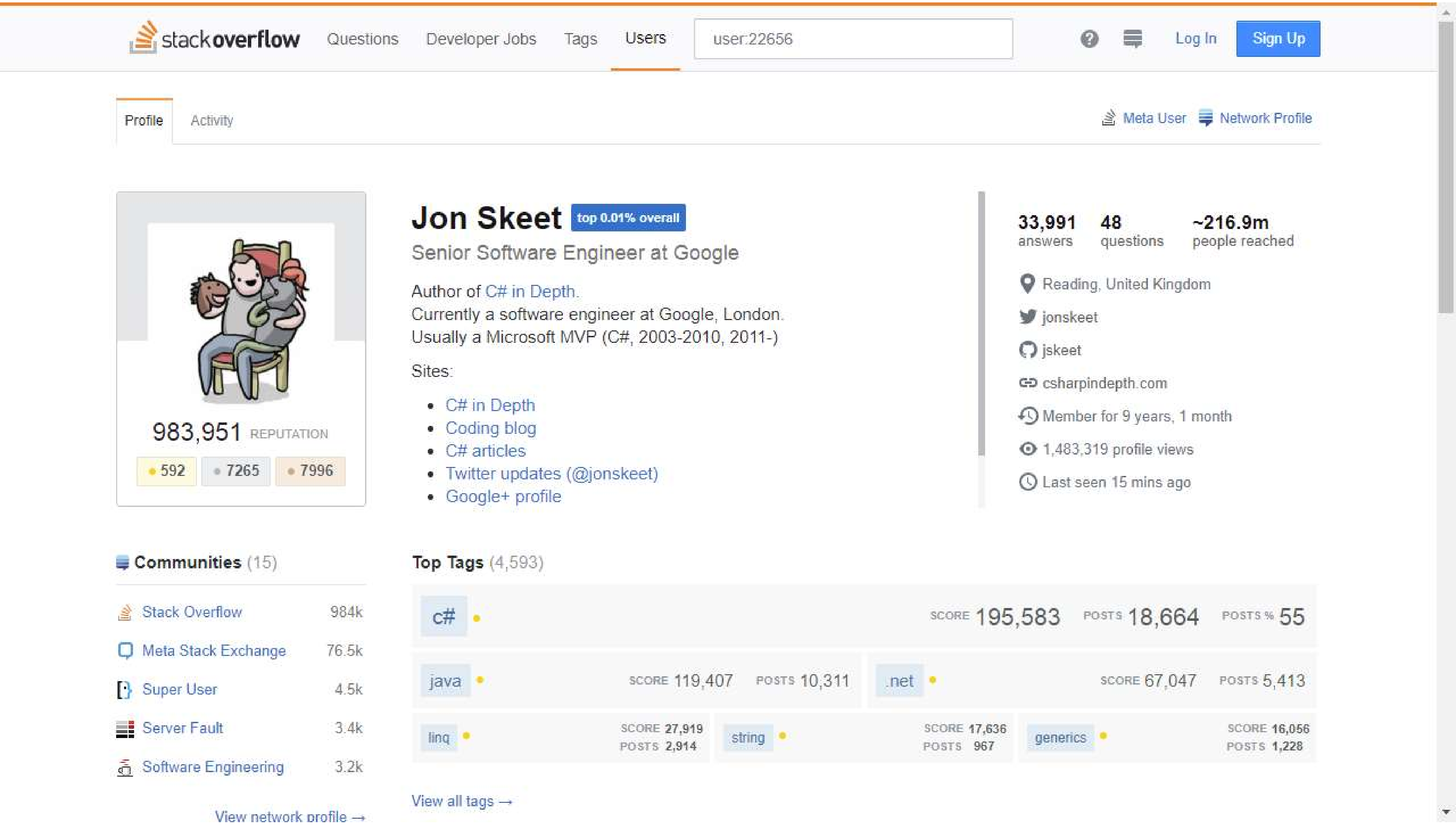
5 votes
2

Q: Apache Spark vs Apache Spark 2 [closed]

What are the improvements **Apache** Spark2 brings compared to **Apache Spark**? From architecture perspective From application point of view or more ...

apache-spark apache-spark-2.0

asked Oct 21 '16 by YoungHobbit



Jon Skeet

top 0.01% overall

Senior Software Engineer at Google

Author of C# in Depth.
Currently a software engineer at Google, London.
Usually a Microsoft MVP (C#, 2003-2010, 2011-)

Sites:

- C# in Depth
- Coding blog
- C# articles
- Twitter updates (@jonskeet)
- Google+ profile

983,951 REPUTATION

● 592 ● 7265 ● 7996

33,991 answers
48 questions
~216.9m people reached

Reading, United Kingdom
jonskeet
jskeet
csharpindepth.com
Member for 9 years, 1 month
1,483,319 profile views
Last seen 15 mins ago

Communities (15)

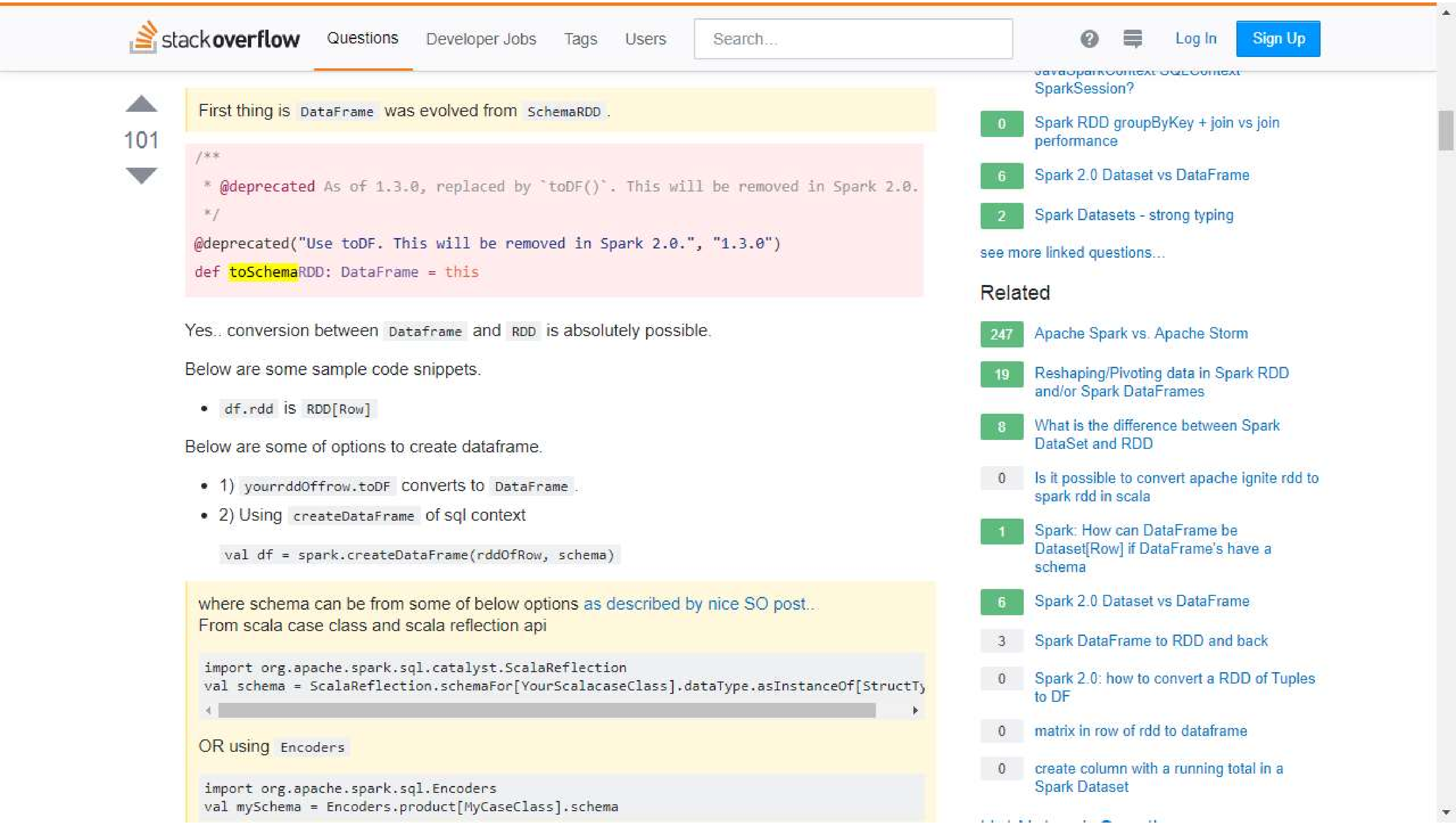
Stack Overflow	984k
Meta Stack Exchange	76.5k
Super User	4.5k
Server Fault	3.4k
Software Engineering	3.2k

View network profile →

Top Tags (4,593)

Tag	Score	Posts	Posts %
c#	195,583	18,664	55
java	119,407	10,311	
.net	67,047	5,413	
linq	27,919	2,914	
string	17,636	967	
generics	16,056	1,228	

View all tags →



101

First thing is `DataFrame` was evolved from `SchemaRDD` .

```
/**
 * @deprecated As of 1.3.0, replaced by `toDF()`. This will be removed in Spark 2.0.
 */
@deprecated("Use toDF. This will be removed in Spark 2.0.", "1.3.0")
def toSchemaRDD: DataFrame = this
```

Yes.. conversion between `Dataframe` and `RDD` is absolutely possible.

Below are some sample code snippets.

- `df.rdd` is `RDD[Row]`

Below are some of options to create dataframe.

- 1) `yourrddOffrow.toDF` converts to `DataFrame` .
- 2) Using `createDataFrame` of sql context

  `val df = spark.createDataFrame(rddOfRow, schema)`

> where schema can be from some of below options as described by nice SO post..
> From scala case class and scala reflection api
>
> ```
> import org.apache.spark.sql.catalyst.ScalaReflection
> val schema = ScalaReflection.schemaFor[YourScalacaseClass].dataType.asInstanceOf[StructTy
> ```
>
> OR using `Encoders`
>
> ```
> import org.apache.spark.sql.Encoders
> val mySchema = Encoders.product[MyCaseClass].schema
> ```

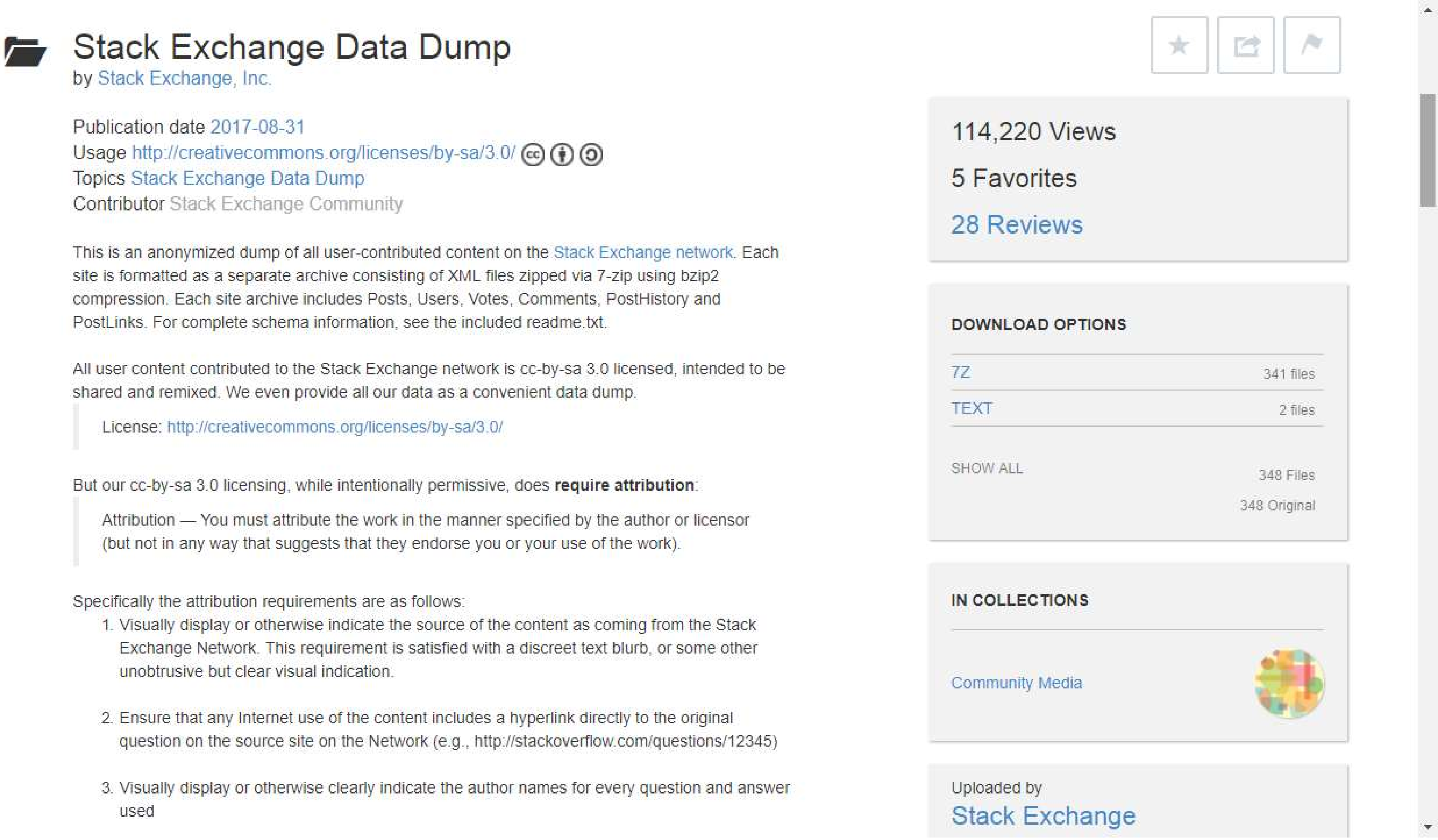
JavaSparkContext SQLContext SparkSession?

0	Spark RDD groupByKey + join vs join performance
6	Spark 2.0 Dataset vs DataFrame
2	Spark Datasets - strong typing

see more linked questions...

Related

247	Apache Spark vs. Apache Storm
19	Reshaping/Pivoting data in Spark RDD and/or Spark DataFrames
8	What is the difference between Spark DataSet and RDD
0	Is it possible to convert apache ignite rdd to spark rdd in scala
1	Spark: How can DataFrame be Dataset[Row] if DataFrame's have a schema
6	Spark 2.0 Dataset vs DataFrame
3	Spark DataFrame to RDD and back
0	Spark 2.0: how to convert a RDD of Tuples to DF
0	matrix in row of rdd to dataframe
0	create column with a running total in a Spark Dataset



Stack Exchange Data Dump

by Stack Exchange, Inc.

Publication date 2017-08-31
Usage http://creativecommons.org/licenses/by-sa/3.0/
Topics Stack Exchange Data Dump
Contributor Stack Exchange Community

This is an anonymized dump of all user-contributed content on the Stack Exchange network. Each site is formatted as a separate archive consisting of XML files zipped via 7-zip using bzip2 compression. Each site archive includes Posts, Users, Votes, Comments, PostHistory and PostLinks. For complete schema information, see the included readme.txt.

All user content contributed to the Stack Exchange network is cc-by-sa 3.0 licensed, intended to be shared and remixed. We even provide all our data as a convenient data dump.

> License: http://creativecommons.org/licenses/by-sa/3.0/

But our cc-by-sa 3.0 licensing, while intentionally permissive, does **require attribution**:

> Attribution — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Specifically the attribution requirements are as follows:

1. Visually display or otherwise indicate the source of the content as coming from the Stack Exchange Network. This requirement is satisfied with a discreet text blurb, or some other unobtrusive but clear visual indication.

2. Ensure that any Internet use of the content includes a hyperlink directly to the original question on the source site on the Network (e.g., http://stackoverflow.com/questions/12345)

3. Visually display or otherwise clearly indicate the author names for every question and answer used

114,220 Views

5 Favorites

28 Reviews

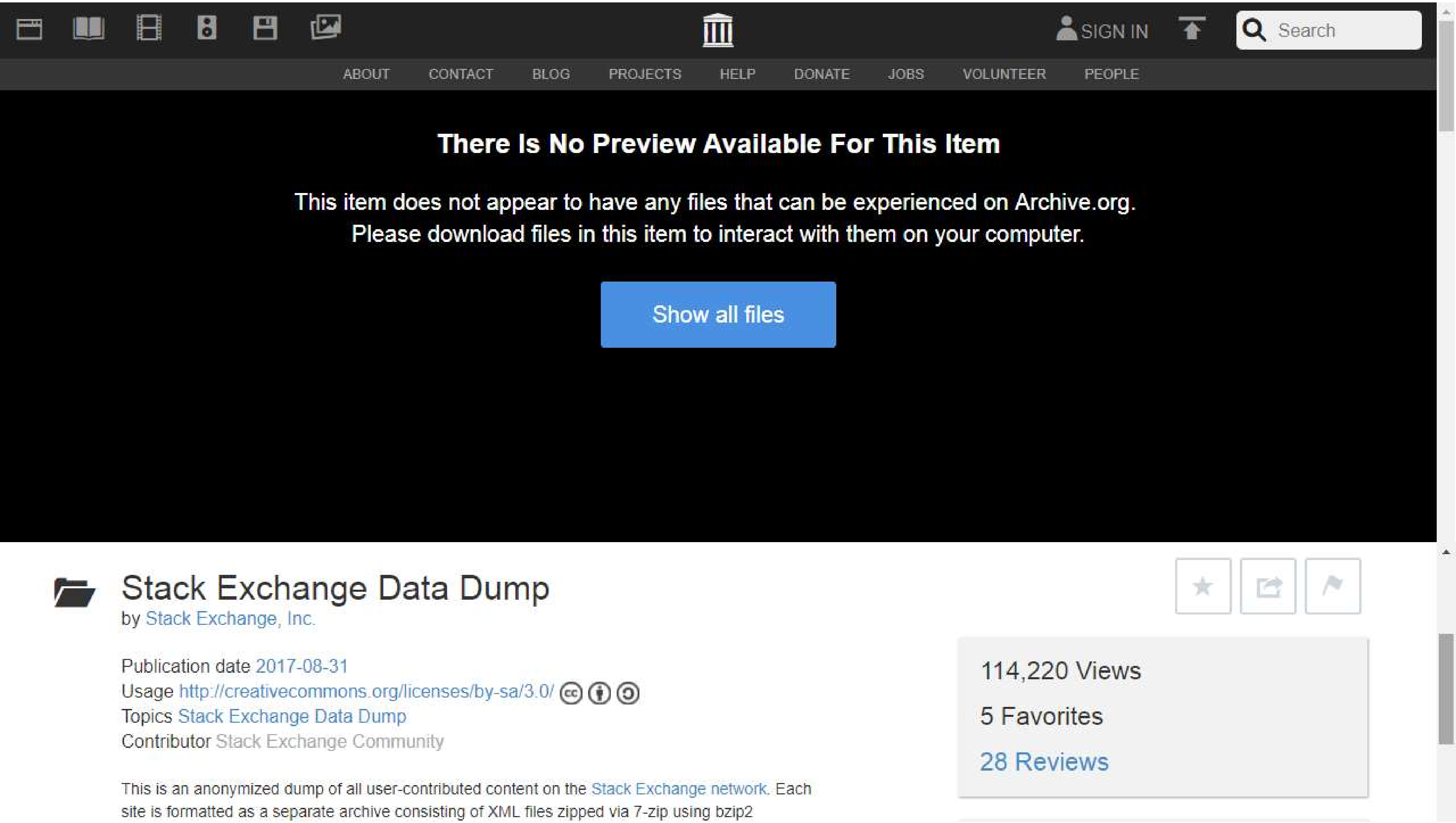
DOWNLOAD OPTIONS

7Z	341 files
TEXT	2 files
SHOW ALL	348 Files 348 Original

IN COLLECTIONS

Community Media

Uploaded by
Stack Exchange



There Is No Preview Available For This Item

This item does not appear to have any files that can be experienced on Archive.org.
Please download files in this item to interact with them on your computer.

Stack Exchange Data Dump

by Stack Exchange, Inc.

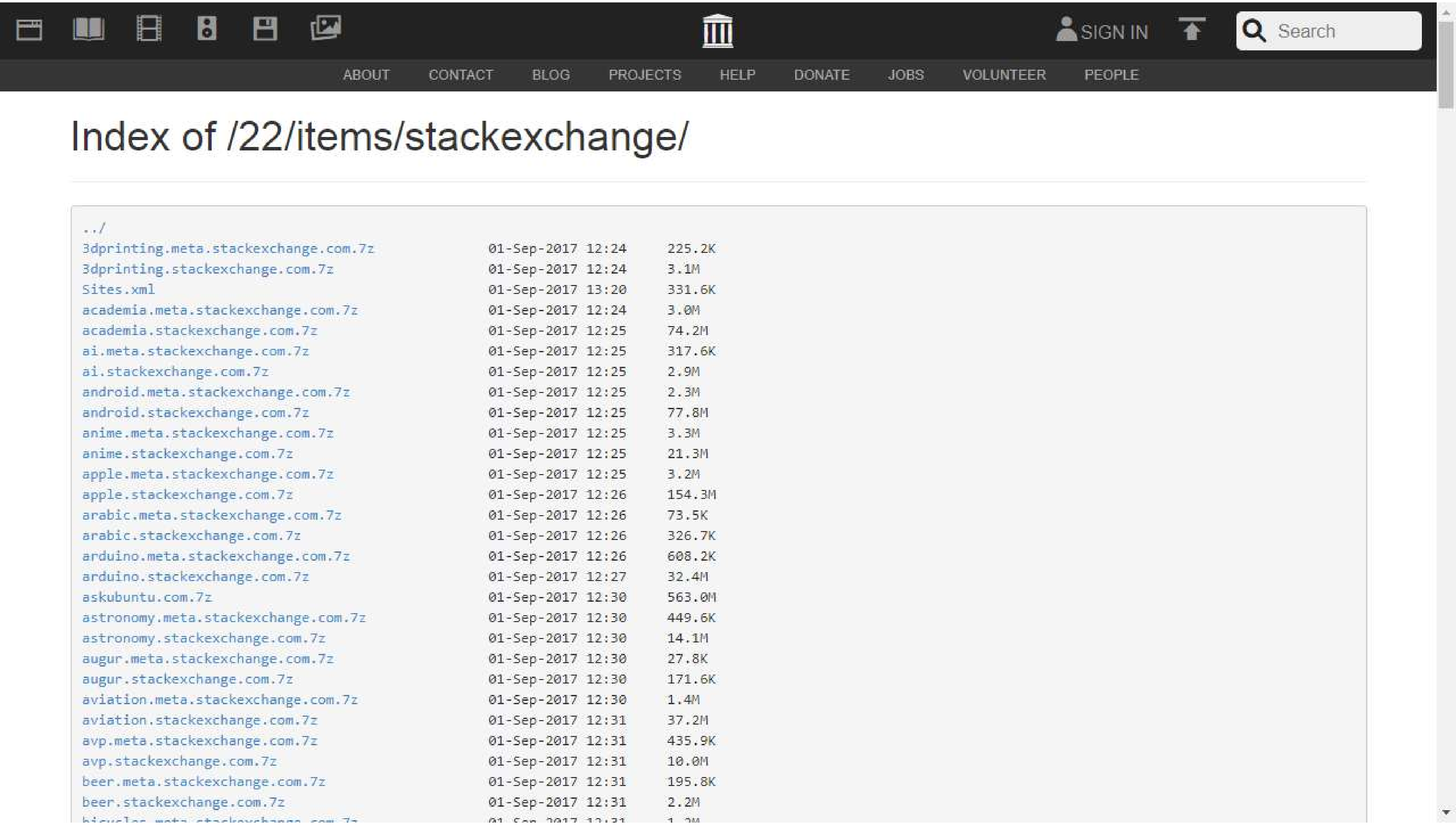
Publication date 2017-08-31
Usage http://creativecommons.org/licenses/by-sa/3.0/
Topics Stack Exchange Data Dump
Contributor Stack Exchange Community

114,220 Views

5 Favorites

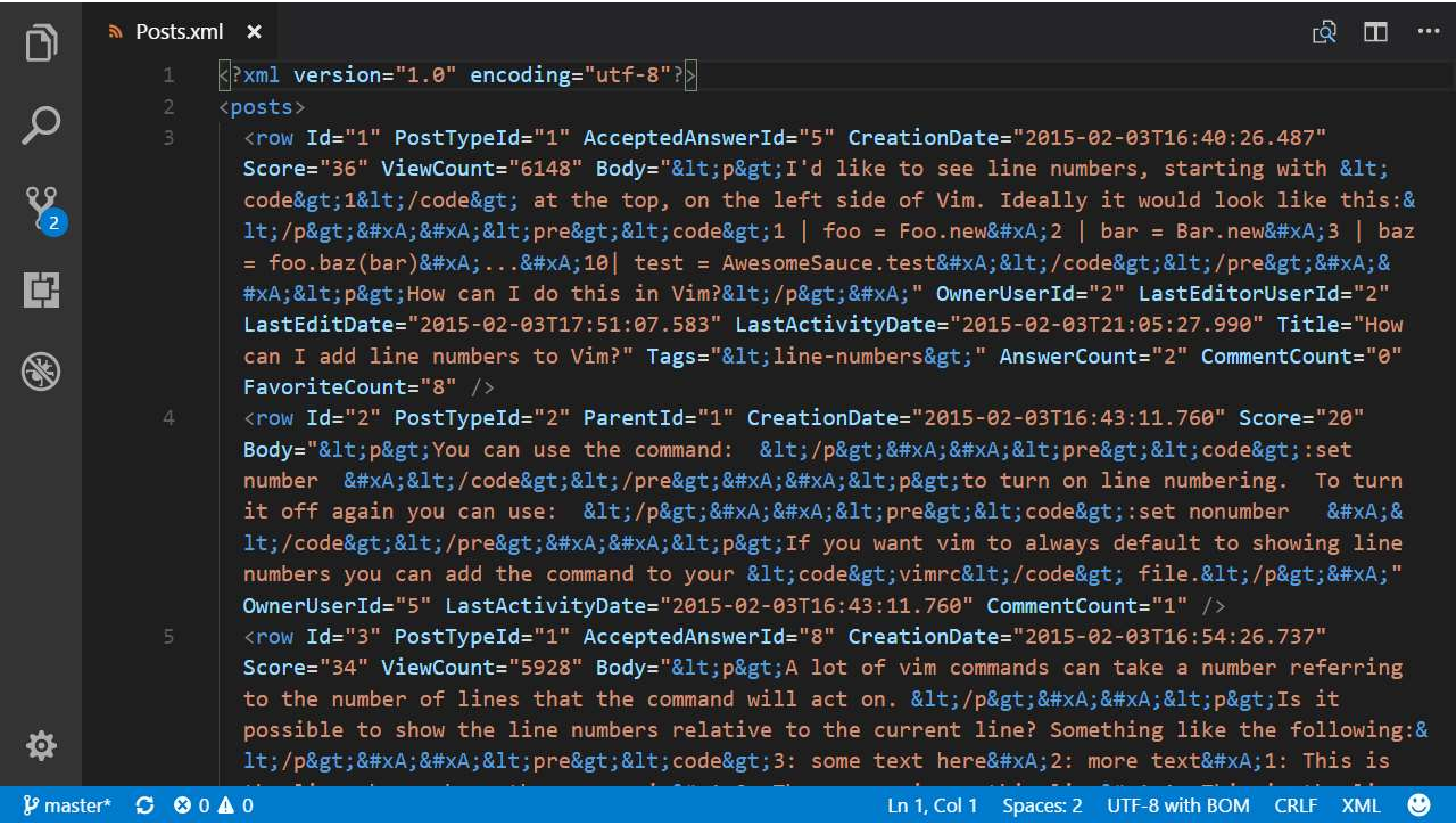
28 Reviews

This is an anonymized dump of all user-contributed content on the Stack Exchange network. Each site is formatted as a separate archive consisting of XML files zipped via 7-zip using bzip2



Index of /22/items/stackexchange/

```
../
3dprinting.meta.stackexchange.com.7z              01-Sep-2017 12:24     225.2K
3dprinting.stackexchange.com.7z                   01-Sep-2017 12:24     3.1M
Sites.xml                                         01-Sep-2017 13:20     331.6K
academia.meta.stackexchange.com.7z                01-Sep-2017 12:24     3.0M
academia.stackexchange.com.7z                     01-Sep-2017 12:25     74.2M
ai.meta.stackexchange.com.7z                      01-Sep-2017 12:25     317.6K
ai.stackexchange.com.7z                           01-Sep-2017 12:25     2.9M
android.meta.stackexchange.com.7z                 01-Sep-2017 12:25     2.3M
android.stackexchange.com.7z                      01-Sep-2017 12:25     77.8M
anime.meta.stackexchange.com.7z                   01-Sep-2017 12:25     3.3M
anime.stackexchange.com.7z                        01-Sep-2017 12:25     21.3M
apple.meta.stackexchange.com.7z                   01-Sep-2017 12:25     3.2M
apple.stackexchange.com.7z                        01-Sep-2017 12:26     154.3M
arabic.meta.stackexchange.com.7z                  01-Sep-2017 12:26     73.5K
arabic.stackexchange.com.7z                       01-Sep-2017 12:26     326.7K
arduino.meta.stackexchange.com.7z                 01-Sep-2017 12:26     608.2K
arduino.stackexchange.com.7z                      01-Sep-2017 12:27     32.4M
askubuntu.com.7z                                  01-Sep-2017 12:30     563.0M
astronomy.meta.stackexchange.com.7z               01-Sep-2017 12:30     449.6K
astronomy.stackexchange.com.7z                    01-Sep-2017 12:30     14.1M
augur.meta.stackexchange.com.7z                   01-Sep-2017 12:30     27.8K
augur.stackexchange.com.7z                        01-Sep-2017 12:30     171.6K
aviation.meta.stackexchange.com.7z                01-Sep-2017 12:30     1.4M
aviation.stackexchange.com.7z                     01-Sep-2017 12:31     37.2M
avp.meta.stackexchange.com.7z                     01-Sep-2017 12:31     435.9K
avp.stackexchange.com.7z                          01-Sep-2017 12:31     10.0M
beer.meta.stackexchange.com.7z                    01-Sep-2017 12:31     195.8K
beer.stackexchange.com.7z                         01-Sep-2017 12:31     2.2M
bicycles.meta.stackexchange.com.7z                01-Sep-2017 12:31     1.2M
```



```xml
<?xml version="1.0" encoding="utf-8"?>
<posts>
  <row Id="1" PostTypeId="1" AcceptedAnswerId="5" CreationDate="2015-02-03T16:40:26.487"
  Score="36" ViewCount="6148" Body="&lt;p&gt;I'd like to see line numbers, starting with &lt;
  code&gt;1&lt;/code&gt; at the top, on the left side of Vim. Ideally it would look like this:&
  lt;/p&gt;&#xA;&#xA;&lt;pre&gt;&lt;code&gt;1 | foo = Foo.new&#xA;2 | bar = Bar.new&#xA;3 | baz
  = foo.baz(bar)&#xA;...&#xA;10| test = AwesomeSauce.test&#xA;&lt;/code&gt;&lt;/pre&gt;&#xA;&
  #xA;&lt;p&gt;How can I do this in Vim?&lt;/p&gt;&#xA;" OwnerUserId="2" LastEditorUserId="2"
  LastEditDate="2015-02-03T17:51:07.583" LastActivityDate="2015-02-03T21:05:27.990" Title="How
  can I add line numbers to Vim?" Tags="&lt;line-numbers&gt;" AnswerCount="2" CommentCount="0"
  FavoriteCount="8" />
  <row Id="2" PostTypeId="2" ParentId="1" CreationDate="2015-02-03T16:43:11.760" Score="20"
  Body="&lt;p&gt;You can use the command:  &lt;/p&gt;&#xA;&#xA;&lt;pre&gt;&lt;code&gt;:set
  number  &#xA;&lt;/code&gt;&lt;/pre&gt;&#xA;&#xA;&lt;p&gt;to turn on line numbering.  To turn
  it off again you can use:  &lt;/p&gt;&#xA;&#xA;&lt;pre&gt;&lt;code&gt;:set nonumber   &#xA;&
  lt;/code&gt;&lt;/pre&gt;&#xA;&#xA;&lt;p&gt;If you want vim to always default to showing line
  numbers you can add the command to your &lt;code&gt;vimrc&lt;/code&gt; file.&lt;/p&gt;&#xA;"
  OwnerUserId="5" LastActivityDate="2015-02-03T16:43:11.760" CommentCount="1" />
  <row Id="3" PostTypeId="1" AcceptedAnswerId="8" CreationDate="2015-02-03T16:54:26.737"
  Score="34" ViewCount="5928" Body="&lt;p&gt;A lot of vim commands can take a number referring
  to the number of lines that the command will act on. &lt;/p&gt;&#xA;&#xA;&lt;p&gt;Is it
  possible to show the line numbers relative to the current line? Something like the following:&
  lt;/p&gt;&#xA;&#xA;&lt;pre&gt;&lt;code&gt;3: some text here&#xA;2: more text&#xA;1: This is
```

Source Data

Users

Posts, History & Links

Tags

Comments

Votes

Badges

Your Hardware Might Vary

Less Resources

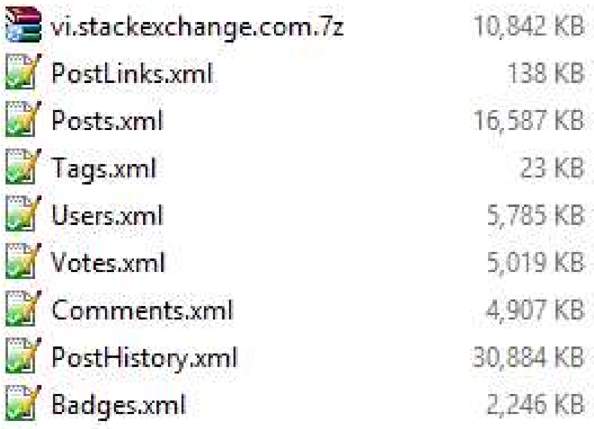
StackExchange

Vi and Vim

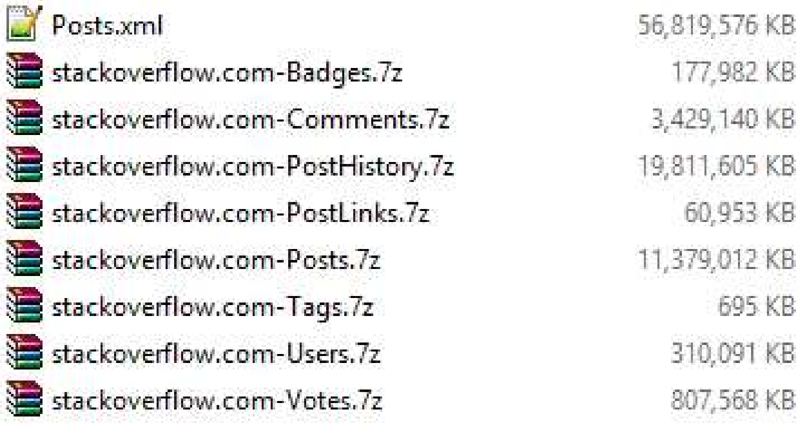
More Resources

stackoverflow

Select What Works for You



vi.stackexchange.com.7z	10,842 KB
PostLinks.xml	138 KB
Posts.xml	16,587 KB
Tags.xml	23 KB
Users.xml	5,785 KB
Votes.xml	5,019 KB
Comments.xml	4,907 KB
PostHistory.xml	30,884 KB
Badges.xml	2,246 KB



Posts.xml	56,819,576 KB
stackoverflow.com-Badges.7z	177,982 KB
stackoverflow.com-Comments.7z	3,429,140 KB
stackoverflow.com-PostHistory.7z	19,811,605 KB
stackoverflow.com-PostLinks.7z	60,953 KB
stackoverflow.com-Posts.7z	11,379,012 KB
stackoverflow.com-Tags.7z	695 KB
stackoverflow.com-Users.7z	310,091 KB
stackoverflow.com-Votes.7z	807,568 KB

Less Resources

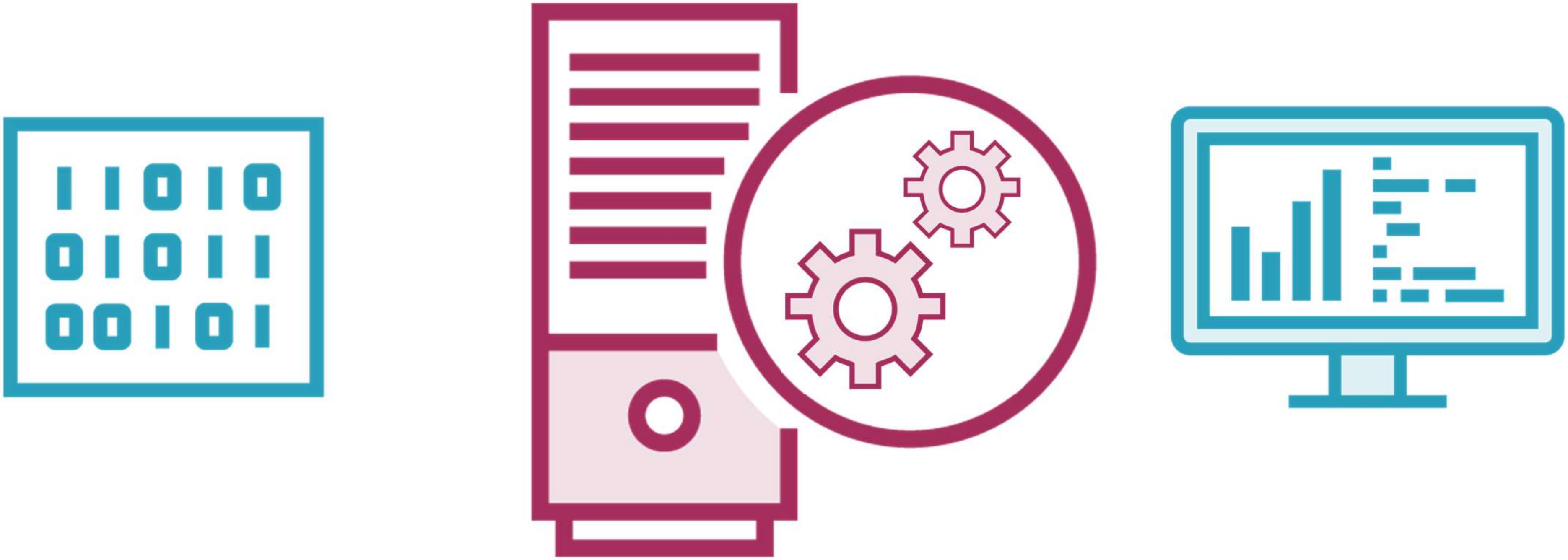
StackExchange

Vi and Vim

More Resources

stackoverflow

Preparing Your Big Data

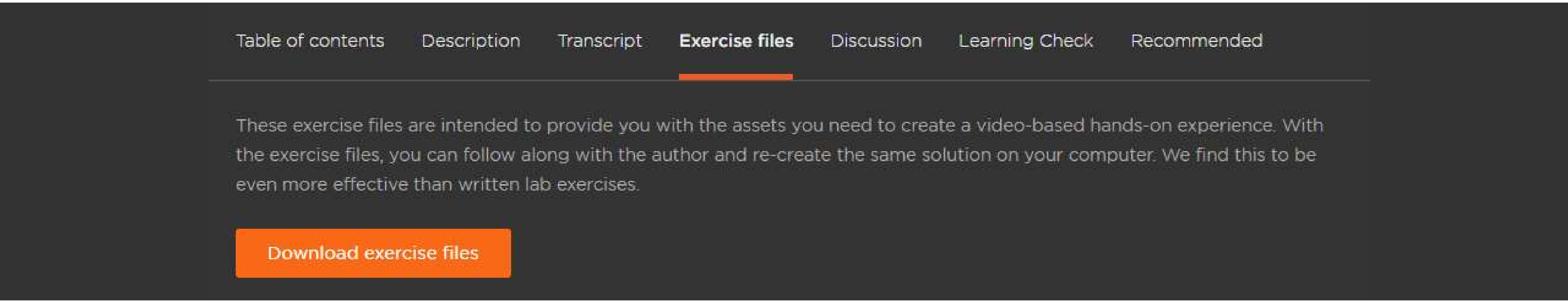


```
cd posts

sbt package

spark2-submit --class "PreparePostsCSVApp"
              target/scala-2.11/posts-project_2.11-1.0.jar
```

Convert Posts.xml to CSV

Data preparation step

Demo Files



These exercise files are intended to provide you with the assets you need to create a video-based hands-on experience. With the exercise files, you can follow along with the author and re-create the same solution on your computer. We find this to be even more effective than written lab exercises.

Takeaway

Environment + Data

- Cloudera cluster
- StackExchange/StackOverflow

Environment: Prereq & known issues

- CDH 5.8.3 and Cloudera Manager 5.8
  - Overview of cluster upgrade
- JDK 1.8
  - Several ways of getting JVM

Takeaway

Installing Spark

- Spark on Yarn 1.6
- Standalone 2.2.0
- Spark on Yarn 2.2

Spark Shell

- Scala

Takeaway

Data

- StackOverflow
- StackExchange

Easily manage with HUE

Prepare data with our first Spark app

Takeaway

Scala