



专注于商业智能BI和大数据的垂直社区平台

# Spark初级课程

讲师：平常心

# 目录



HDP的Spark安装

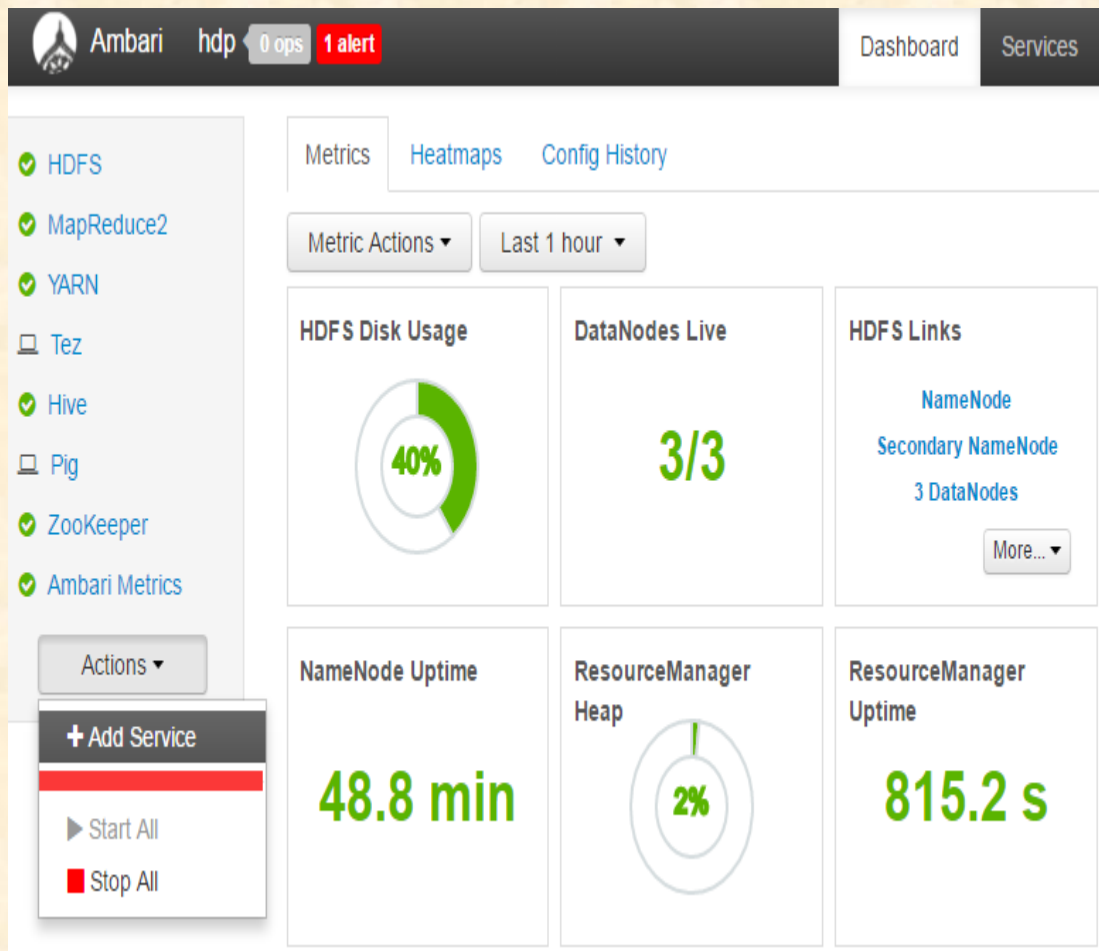


CDH的Spark安装



HDP和CDH对比

# HDP安装Spark服务



## Add Service Wizard

<input type="checkbox"/>	Falcon	0.6.1.2.4	Data management and processing platform
<input type="checkbox"/>	Storm	0.10.0.2.4	Apache Hadoop Stream processing framework
<input type="checkbox"/>	Flume	1.5.2.2.4	A distributed service for collecting, aggregating, and moving large amounts of streaming data into HDFS
<input type="checkbox"/>	Accumulo	1.7.0.2.4	Robust, scalable, high performance distributed key/value store.
<input checked="" type="checkbox"/>	Ambari Metrics	0.1.0	A system for metrics collection that provides storage and retrieval capability for metrics collected from the cluster
<input type="checkbox"/>	Atlas	0.5.0.2.4	Atlas Metadata and Governance platform
<input type="checkbox"/>	Kafka	0.9.0.2.4	A high-throughput distributed messaging system
<input type="checkbox"/>	Knox	0.6.0.2.4	Provides a single point of authentication and access for Apache Hadoop services in a cluster
<input type="checkbox"/>	Mahout	0.9.0.2.4	Project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification
<input type="checkbox"/>	Ranger	0.5.0.2.4	Comprehensive security for Hadoop
<input type="checkbox"/>	Ranger KMS	0.5.0.2.4	Key Management Server
<input type="checkbox"/>	Slider	0.80.0.2.4	A framework for deploying, managing and monitoring existing distributed applications on YARN.
<input type="checkbox"/>	SmartSense	1.2.1.0-70	SmartSense - Hortonworks SmartSense Tool (HST) helps quickly gather configuration, metrics, logs from common HDP services that aids to quickly troubleshoot support cases and receive cluster-specific recommendations.
<input checked="" type="checkbox"/>	Spark	1.6.0.2.4	Apache Spark is a fast and general engine for large-scale data processing.

Next →

# CDH安装Spark服务

clouderaMANAGER

群集 ▾ 主机 ▾ 诊断 ▾ 审核 图表 ▾ 管理 ▾

主页

状态 所有运行状况问题 配置 5 所有最新命令

Cluster 1 (CDH 5.8.0, Parcel)

主机

HDFS

Hive

Hue

Oozie

YARN (MR2 Inclu...

ZooKeeper

添加服务

启动

停止

重启

滚动重启

部署客户端配置

部署 Kerberos 客户端配置

升级群集

刷新群集

刷新动态资源池

Inspect Hosts in Cluster

启用 Kerberos

设置 HDFS 静态数据加密

查看客户端配置 URL

重命名群集

进入维护模式

视图维护模式状态

图表

群集 CPU

Cluster 1, 整个主机中的主机 CPU 使用率 47.1%

	Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log. Before adding this service, ensure that either the Kafka parcel is activated or the Kafka package is installed.
	Key-Value Store Indexer	键值 Store Indexer 侦听 HBase 中所含表内的数据变化, 并使用 Solr 为其创建索引。
	MapReduce	Apache Hadoop MapReduce 支持对整个群集中的大型数据集进行分布式计算(需要 HDFS)。建议改用 YARN (包括 MapReduce 2)。包括 MapReduce 用于向后兼容性。
	Oozie	Oozie 是群集中管理数据处理作业的工作流协调服务。
	Sentry	Sentry 服务存储身份验证政策元数据并为客户端提供对该元数据的并发安全访问。
	Solr	Solr 是一个分布式服务, 用于编制存储在 HDFS 中的数据的索引并搜索这些数据。
	Spark	Apache Spark is an open source cluster computing system. This service runs Spark as an application on YARN.
	Spark (Standalone)	Apache Spark is an open source cluster computing system. This is the standalone version of the service which does not use YARN for resource management. Cloudera recommends using Spark on YARN instead of this standalone version.
	Sqoop 1 Client	Configuration and connector management for Sqoop 1.
	Sqoop 2	Sqoop 是一个设计用于在 Apache Hadoop 和结构化数据存储(如关系数据库)之间高效地传输大批量数据的工具。Cloudera Manager 支持的版本为 Sqoop 2。
	YARN (MR2 Included)	Apache Hadoop MapReduce 2.0 (MRv2) 或 YARN 是支持 MapReduce 应用程序的数据计算框架(需要 HDFS)。
	ZooKeeper	Apache ZooKeeper 是用于维护和同步配置数据的集中服务。

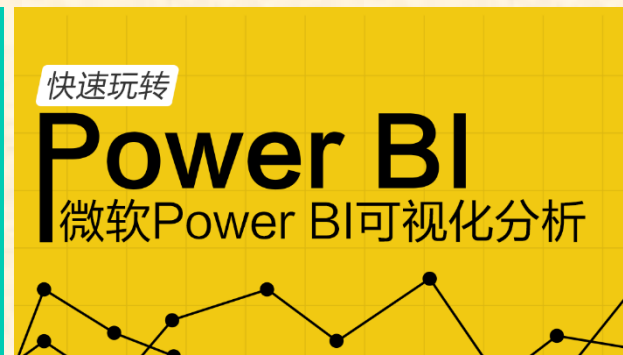
返回 继续



# 对比说明

	HDP	CDH
运营管理界面	Ambari 目前汉化自己开发	CM 支持国际化
Hive	hive.execution.engine支持 MR和Tez	hive.execution.engine支持 MR和Spark
查询引擎	Hive on Tez Phoenix Hawq	Hive on Spark Impala
Spark	Yarn模式	Yarn和Standalone 不支持spark-sql
组件	开源组件 Falcon Atlas	有自主开发组件 Impala <b>Navigator</b> 免费版本 企业版本

更多商业智能BI和大数据精品视频尽在 [www.hellobi.com](http://www.hellobi.com)



BI、商业智能  
数据挖掘 大数据  
数据分析  
R Python  
机器学习  
Tableau  
QLIKVIEW  
Hive Hadoop  
BIWORK  
BAO胖子 seng  
曹浩 贝克汉姆