Hadoop 家族学习路线图

主要介绍 Hadoop 家族产品，常用的项目包括 Hadoop, Hive, Pig, HBase, Sqoop, Mahout, Zookeeper, Avro, Ambari, Chukwa，新增加的项目包括，YARN, Hcatalog, Oozie, Cassandra, Hama, Whirr, Flume, Bigtop, Crunch, Hue 等。

从 2011 年开始，中国进入**大数据**风起云涌的时代，以 **Hadoop** 为代表的家族软件，占据了大数据处理的广阔地盘。开源界及厂商，所有数据软件，无一不向 Hadoop 靠拢。Hadoop 也从小众的高富帅领域，变成了大数据开发的标准。在 Hadoop 原有技术基础之上，出现了 Hadoop 家族产品，通过“大数据”概念不断创新，推出科技进步。

作为 IT 界的开发人员，我们也要跟上节奏，抓住机遇，跟着 Hadoop 一起雄起！

前言

使用 Hadoop 已经有一段时间了，从开始的迷茫，到各种的尝试，到现在组合应用....慢慢地涉及到数据处理的事情，已经离不开 hadoop 了。Hadoop 在大数据领域的成功，更引发了它本身的加速发展。现在 Hadoop 家族产品，已经达到 20 个了之多。

有必要对自己的知识做一个整理了，把产品和技术都串起来。不仅能加深印象，更可以对以后的技术方向，技术选型做好基础准备。

本文为“Hadoop 家族”开篇，Hadoop 家族学习路线图

目录

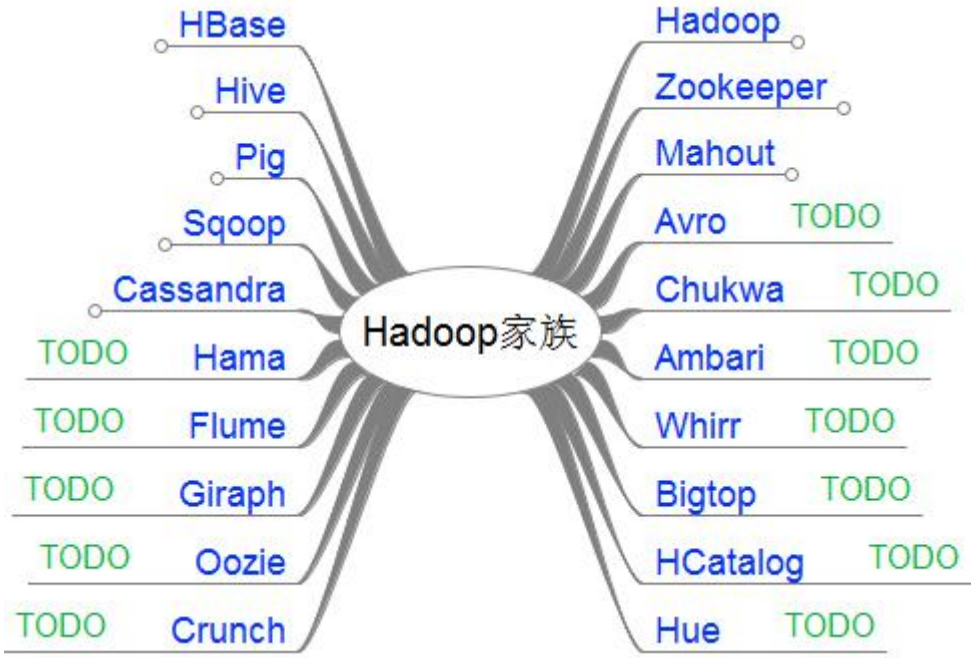
1. Hadoop 家族产品
2. Hadoop 家族学习路线图

1. Hadoop 家族产品

截止到 2013 年，根据 cloudera 的统计，Hadoop 家族产品已经达到 20 个！
http://blog.cloudera.com/blog/2013/01/apache-hadoop-in-2013-the-state-of-the-platform/

接下来，我把这 20 个产品，分成了 2 类。

- 第一类，是我已经掌握的
- 第二类，是 TODO 准备继续学习的



一句话产品介绍:

- Apache Hadoop: 是 Apache 开源组织的一个分布式计算开源框架，提供了一个分布式文件系统子项目(HDFS)和支持 MapReduce 分布式计算的软件架构。

- Apache Hive: 是基于 Hadoop 的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，通过类 SQL 语句快速实现简单的 MapReduce 统计，不必开发专门的 MapReduce 应用，十分适合数据仓库的统计分析。

- Apache Pig: 是一个基于 Hadoop 的大规模数据分析工具，它提供的 SQL-LIKE 语言叫 Pig Latin，该语言的编译器会把类 SQL 的数据分析请求转换为一系列经过优化处理的 MapReduce 运算。

- Apache HBase: 是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，利用 HBase 技术可在廉价 PC Server 上搭建起大规模结构化存储集群。

- Apache Sqoop: 是一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具，可以将一个关系型数据库（MySQL ,Oracle ,Postgres 等）中的数据导进到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导进到关系型数据库中。

- Apache Zookeeper: 是一个为分布式应用所设计的分布的、开源的协调服务，它主要是用来解决分布式应用中经常遇到的一些数据管理问题，简化分布式应用协调及其管理的难度，提供高性能的分布式服务

- Apache Mahout:是基于Hadoop的机器学习和数据挖掘的一个分布式框架。Mahout用 MapReduce 实现了部分数据挖掘算法，解决了并行挖掘的问题。

- Apache Cassandra:是一套开源分布式 NoSQL 数据库系统。它最初由 Facebook 开发，用于储存简单格式数据，集 Google BigTable 的数据模型与 Amazon Dynamo 的完全分布式的架构于一身

- Apache Avro: 是一个数据序列化系统，设计用于支持数据密集型，大批量数据交换的应用。Avro 是新的数据序列化格式与传输工具，将逐步取代 Hadoop 原有的 IPC 机制

- Apache Ambari: 是一种基于 Web 的工具，支持 Hadoop 集群的供应、管理和监控。

- Apache Chukwa: 是一个开源的用于监控大型分布式系统的数据收集系统，它可以将各种各样类型的数据收集成适合 Hadoop 处理的文件保存在 HDFS 中供 Hadoop 进行各种 MapReduce 操作。

- Apache Hama: 是一个基于 HDFS 的 BSP（Bulk Synchronous Parallel)并行计算框架, Hama 可用于包括图、矩阵和网络算法在内的大规模、大数据计算。

- Apache Flume: 是一个分布的、可靠的、高可用的海量日志聚合的系统，可用于日志数据收集，日志数据处理，日志数据传输。

- Apache Giraph: 是一个可伸缩的分布式迭代图处理系统， 基于 Hadoop 平台，灵感来自 BSP (bulk synchronous parallel) 和 Google 的 Pregel。

- Apache Oozie: 是一个工作流引擎服务器，用于管理和协调运行在 Hadoop 平台上（HDFS、Pig 和 MapReduce）的任务。

- Apache Crunch: 是基于 Google 的 FlumeJava 库编写的 Java 库，用于创建 MapReduce 程序。与 Hive，Pig 类似，Crunch 提供了用于实现如连接数据、执行聚合和排序记录等常见任务的模式库

- Apache Whirr: 是一套运行于云服务的类库（包括 Hadoop），可提供高度的互补性。Whirr 学支持 Amazon EC2 和 Rackspace 的服务。
- Apache Bigtop: 是一个对 Hadoop 及其周边生态进行打包，分发和测试的工具。
- Apache HCatalog: 是基于 Hadoop 的数据表和存储管理，实现中央的元数据和模式管理，跨越 Hadoop 和 RDBMS，利用 Pig 和 Hive 提供关系视图。
- Cloudera Hue: 是一个基于 WEB 的监控和管理系统，实现对 HDFS，MapReduce/YARN, HBase, Hive, Pig 的 web 化操作和管理。

2. Hadoop 家族学习路线图

下面我将分别介绍各个产品的安装和使用，以我经验总结我的学习路线。

Hadoop

- Hadoop 学习路线图
- Yarn 学习路线图
- 用 Maven 构建 Hadoop 项目
- Hadoop 历史版本安装
- Hadoop 编程调用 HDFS
- 海量 Web 日志分析 用 Hadoop 提取 KPI 统计指标
- 用 Hadoop 构建电影推荐系统
- 创建 Hadoop 母体虚拟机
- 克隆虚拟机增加 Hadoop 节点
- R 语言为 Hadoop 注入统计血脉
- RHadoop 实践系列之一 Hadoop 环境搭建

Hive

- Hive 学习路线图
- Hive 安装及使用攻略
- Hive 导入 10G 数据的测试
- R 利剑 NoSQL 系列文章 之 Hive
- 用 RHive 从历史数据中提取逆回购信息

Pig

- Pig 学习路线图

Zookeeper

- Zookeeper 学习路线图
- ZooKeeper 伪分步式集群安装及使用
- ZooKeeper 实现分布式队列 Queue
- ZooKeeper 实现分布式 FIFO 队列

Hbase

- HBase 学习路线图
- RHadoop 实践系列之四　rhbase 安装与使用

Mahout

- Mahout 学习路线图
- 用 R 解析 Mahout 用户推荐协同过滤算法(UserCF)
- RHadoop 实践系列之三 R 实现 MapReduce 的协同过滤算法

- 用 Maven 构建 Mahout 项目
- Mahout 推荐算法 API 详解
- 从源代码剖析 Mahout 推荐引擎
- Mahout 分步式程序开发 基于物品的协同过滤 ItemCF
- Mahout 分步式程序开发 聚类 Kmeans
- 用 Mahout 构建职位推荐引擎

Sqoop

- Sqoop 学习路线图

Cassandra

- Cassandra 学习路线图
- Cassandra 单集群实验 2 个节点
- R 利剑 NoSQL 系列文章 之 Cassandra