



# DataDiscovery 数据发 现

## SDC Hadoop

---

### 产品概述

文档版本 01 发布日期 2016-08-12

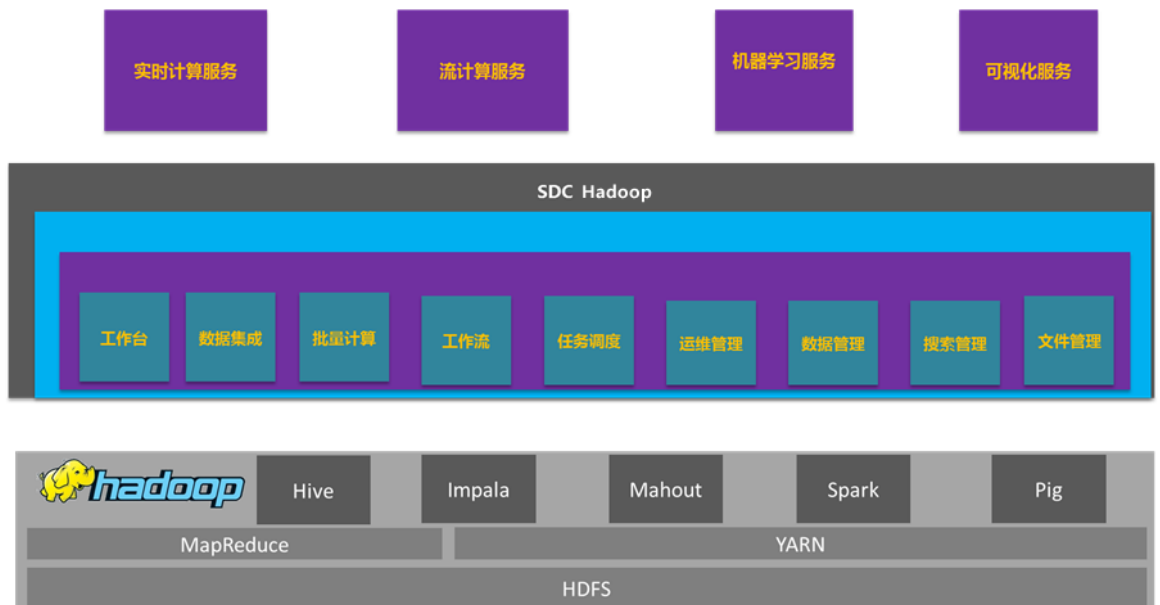
---





# 1 概述

当前，大数据已然成为所有行业最具价值的潜在财富，政府、企业都已经意识从数据提炼知识的重要性。面对这些持续激增的海量数据信息，人类对数据的管理挖掘和需求越来越旺盛，在解决问题的时候，他所牵扯到的数据量和维度也是不断爆炸，不断增长，如何更好、更轻松地挖掘和管理数据是大问题。不同于百万级、千万级数的数据分析与挖掘，大数据通常面对的是几十亿行、甚至上百 TB 级超大数据量。数据集市中的数据通常来自不同业务、不同架构的系统中，复杂性极高，进行模型运算时可能是四维的甚至更多条件的计算分析。这不仅要求分析准确，还必须满足客户的应用场景性能要求。



大数据第一件事要解决把数据管理起来，一是从容量上横向扩展，二是在解决不同的数据问题的时候需要不同的数据技术，包括底层的 Hadoop、Spark、Storm 等分布式技术，在管理层面也需要这样的管理框架，要能够即插即用管理各种各样的计算框架。SDC Hadoop 提供基础的分布式文件系统作为存储引擎，YARN 作为资源管理框架，组合了一系列 Apache 项目，为用户提供统一工作台实现数据的采集、存储、数据同步、批处理、工作流分析以及全文搜索功能，四方伟业改进了基于 YARN 资源管理框架，可以在同一个 HDFS 数据集上动态创建流式计算集群、实时分析集群、Map/Reduce 批处理集群以及实时流处理集群，提供多部门间资源配额管理、动态资源调配、资源共享的能力，为企业建立全流程数据开发平台提供支撑。

## 2 应用场景

海量数据搜索

混合计算业务

大数据管理

大数据开发

定制调度策略

大数据运维

## 3 应用案例

已成功服务于包括金融、电信、政府与公共事业、交通、教育等多个行业的客户。

