**大数据项目开发**

**设备需求**

**目录**

[一、平台概述 3](#_Toc414454775)

[1.1 TDH 3](#_Toc414454776)

[1.2 CDH 4](#_Toc414454777)

[1.3 APACHE HADOOP 6](#_Toc414454778)

[二、平台服务器需求 7](#_Toc414454779)

[2.1大数据开发平台服务器 7](#_Toc414454780)

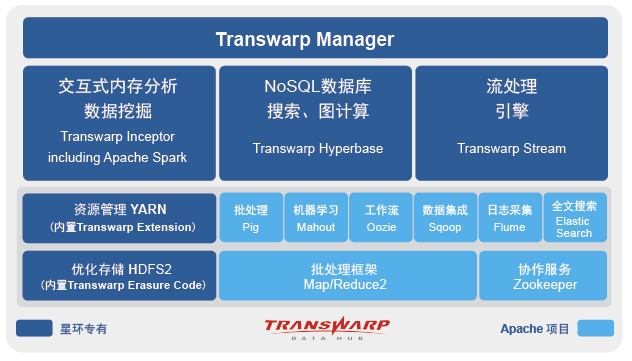
[2.2项目开发及管理服务器 8](#_Toc414454781)

# 一、平台概述

当前大数据分析平台版本有很多，针对公司大数据应用项目开发特点以及为了兼顾产品后期开发和维护等各种需求，公司在搭建大数据开发平台采用三种Hadoop产品，部署在开发平台上，分别是TDH,CDH以及APACHE HADOOP .

## 1.1 TDH

TDH由上海星环科技基于Apache Hadoop为企业开发了一站式大数据综合平台Transwarp Data Hub（简称TDH）， 通过提供从数据存储、分布式计算、数据分析挖掘以及数据可视化的整套支持，为企业建立一个统一的数据和计算平台。企业用户可以基于大数据综合平台上采集、存储、分析、搜索、挖掘海量数据及其内在价值。Transwarp Data Hub 体系架构如下:

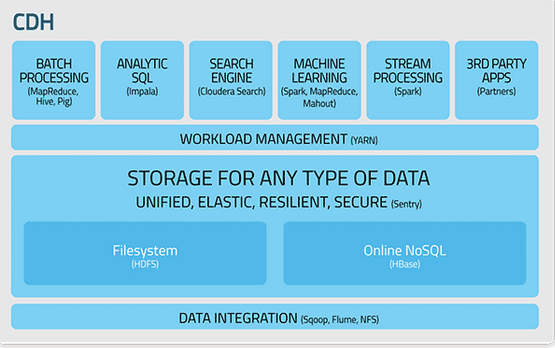


组件介绍如下：

|  |  |
| --- | --- |
| **组件名称** | **功能描述** |
| Transwarp Hadoop | 分布式数据存储及计算。 |
| Transwarp Hyperbase | 针对结构化和非结构化数据的分布式大数据数据库。它可以提供大数据量的高速读写操作。 |
| YARN | 集群资源分配，监控及管理等 |
| Transwarp Inceptor  (include many component) | 提供高速SQL分析和R语言数据挖掘能力 |
| Transwarp Stream | 实时流处理引擎提供了流计算服务 |
| Zookeeper | 分布式协调服务 |
| Pig | 批处理 |
| Mahout | 提供机器学习算法库 |
| Oozie | 工作流 |
| Sqoop | ETL工具 |
| Flume | 日志聚合 |

## 1.2 CDH

CDH是由Cloudera 公司基于Apache Hadoop为企业开发了一站式大数据综合平台Transwarp Data Hub（简称TDH）， 通过提供从数据存储、分布式计算、数据分析挖掘以及数据可视化的整套支持，为企业建立一个统一的数据和计算平台。企业用户可以基于大数据综合平台上采集、存储、分析、搜索、挖掘海量数据及其内在价值。CDH体系架构如下:

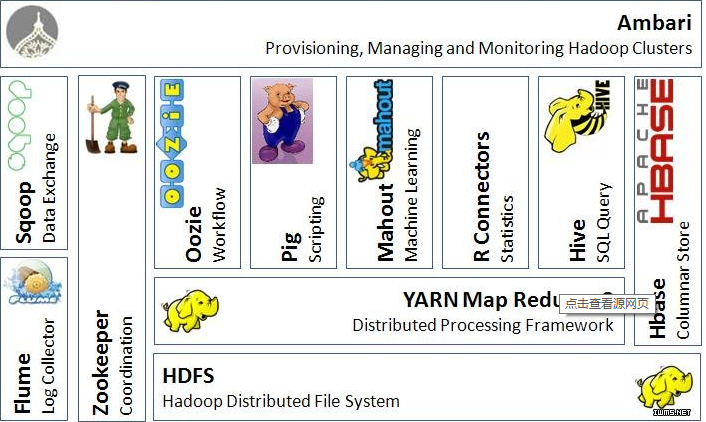


组件介绍如下：

|  |  |
| --- | --- |
| **组件名称** | **功能描述** |
| HDFS | 分布式数据存储 |
| HBASE | 分布式数据库 |
| YARN | 集群资源分配，监控及管理等 |
| Mapredue | 分布式批计算框架 |
| Hive | 大数据分布式数据仓库软件。 |
| Pig | 大数据分布式数据分析语言和运行平台。 |
| Impala | SQL语句解析引擎 |
| spark | 分布式内存计算框架 |
| Spark streaming | 实时流处理引擎提供了流计算服务 |
| Zookeeper | 分布式协调服务 |
| Mahout | 提供机器学习算法库 |
| Oozie | 工作流 |
| Sqoop | ETL工具 |
| Flume | 日志聚合 |

## 1.3 APACHE HADOOP

APACHE HADOOP由开源社区提供的大数据分析平台， 提供从数据存储、分布式计算、数据分析挖掘以及数据可视化的整套支持，为企业建立一个统一的数据和计算平台。企业用户可以基于大数据综合平台上采集、存储、分析、搜索、挖掘海量数据及其内在价值。APACHE HADOOP架构如下：



组件介绍如下：

|  |  |
| --- | --- |
| **组件名称** | **功能描述** |
| HDFS | 分布式数据存储 |
| HBASE | 分布式数据库 |
| YARN | 集群资源分配，监控及管理等 |
| Mapredue | 分布式批计算框架 |
| Hive | 大数据分布式数据仓库软件。 |
| Pig | 大数据分布式数据分析语言和运行平台。 |
| Spark | 分布式内存计算框架 |
| Storm | 实时流处理引擎提供了流计算服务 |
| Zookeeper | 分布式协调服务 |
| Mahout | 提供机器学习算法库 |
| Oozie | 工作流 |
| Sqoop | ETL工具 |
| Flume | 日志聚合 |

# 二、平台服务器需求

## 2.1大数据开发平台服务器

为了满足产品的开发及维护等需求，机器节点及配置如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **集群** | **节点** | **节点数** | **部署组件** | **建议配置** |
| **Hadoop集群**  **+**  **流处理集群** | 元数据管理节点 | 2 | HDFS,HBASE,YARN,  INCEPTOR，SPARK,  STORM, **Kerberos, LDAP,**  **MAPREDUCE及动态扩展其它组件** | 1. 2颗6-8核CPU 2. 64G 以上的内存 3. 2块500G SAS硬盘以上及1T硬盘 4. 千兆以上的以太网卡 |
| 数据存储节点 | 4 | HDFS,HBASE,  YARN,  INCEPTOR，SPARK,  STORM, **Kerberos, LDAP,**  **MAPREDUCE及动态扩展其它组件** | 1、2颗6-8核 CPU  2、48G以上 内存  3、2块500G SAS硬盘以上及3T硬盘  4-千兆以上的以太网卡 |
| 协调服务管理节点 | 3 | ZOOKEEPER,  HDFS**及动态扩展其它组件** | 1、2颗6-8核CPU  2、48G以上 内存  3、2块500G SAS硬盘以上及1T硬盘  4-千兆以上的以太网卡 |

## 2.2项目开发及管理服务器

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **开发服务器** | **节点** | **节点数** | **部署组件** | **备注** |
| 软件开发服务器 | 软件开发服务器 | 1 | 数据库，中间件，集群负载，  MAVEN库SVN,JIRA | 1、2颗6-8核CPU  2、48G以上 内存  3、2块500G SAS硬盘以上及1T硬盘  4-千兆以上的以太网卡 |

注：

1. 合计10台
2. 建议提高物理主机的配置，采用虚机（前提一定要保证虚机的配置资源）
3. 大数据平台服务及开发服务建议跟其它使用的机器分开