基于Hadoop的
海量文本处理系统介绍

江南计算技术研究所

纲要

一、系统简介

- 基于Hadoop系统的文本处理
- 基于Nutch的Framework创建
- 提供Eclipse插件来支持用户开发
- 通过Web方式进行运行维护管理

一、系统简介

- 30台双路四核商用服务器
  - ✓ Intel(R) Xeon(R) CPU E5450 @ 3.00GHz
  - ✓ 16GB内存
  - ✓ 8x500GB SAS硬盘
- 网络：20Gb/s IB网＋千兆以太网

一、系统简介

- 操作系统

  Red Hat Enterprise Linux AS release 4 (Nahant Update 7)

- 内核

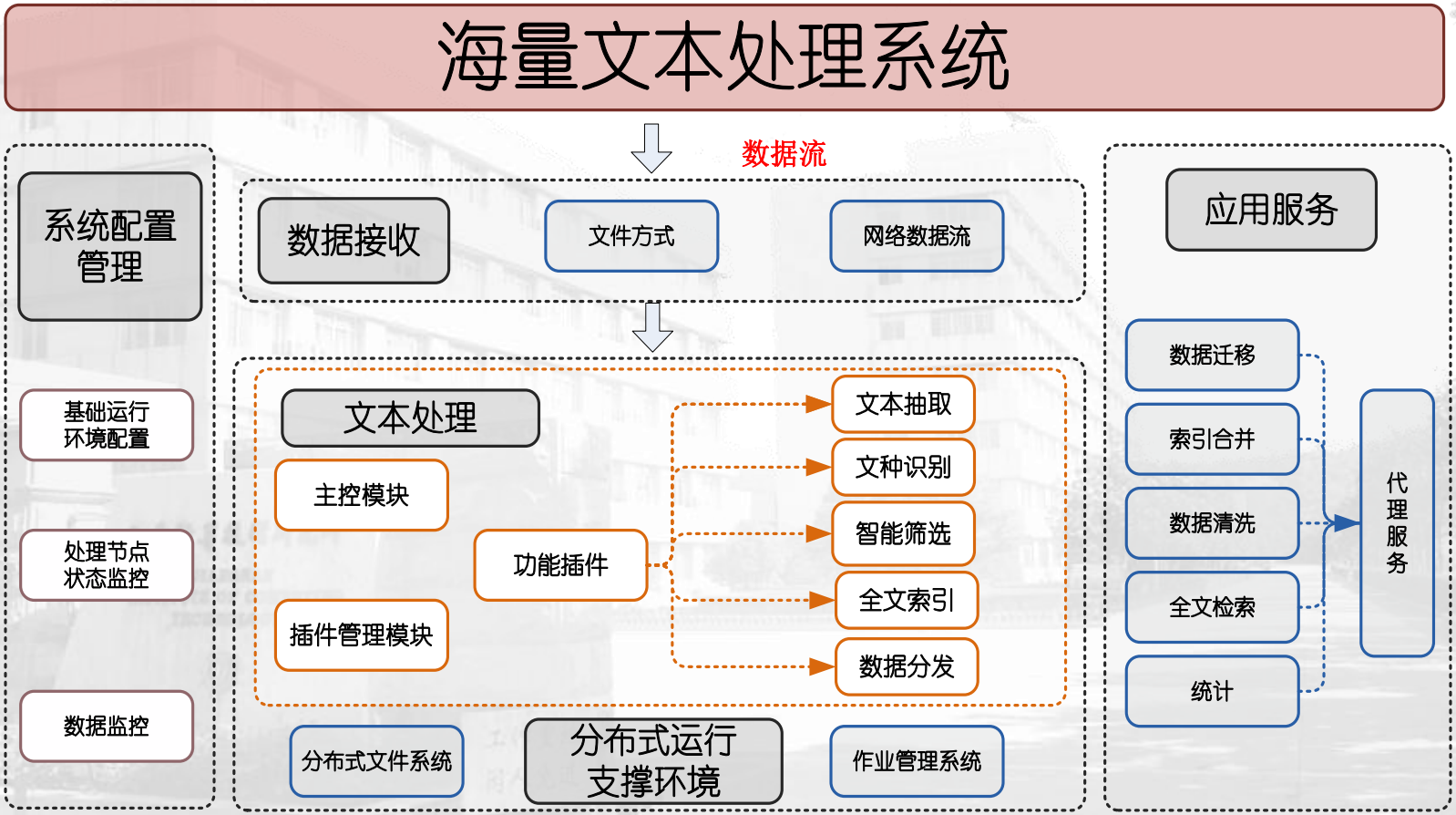
  2.6.9-78.ELsmp SMP x86_64

一、系统简介

- 应用软件
  - ✓ Java(TM) SE Runtime Environment (build 1.6.0_13-b03)
  - ✓ gcc version 3.4.6 20060404 (Red Hat 3.4.6-10)
- Hadoop版本

  hadoop-0.12.2-core-jn.jar（补丁版本）

二、系统组成


海量文本处理系统
数据流
系统配置管理
基础运行环境配置
处理节点状态监控
数据监控
数据接收
文件方式
网络数据流
文本处理
主控模块
功能插件
插件管理模块
文本抽取
文种识别
智能筛选
全文索引
数据分发
分布式文件系统
分布式运行支撑环境
作业管理系统
应用服务
数据迁移
索引合并
数据清洗
全文检索
统计
代理服务

二、系统组成

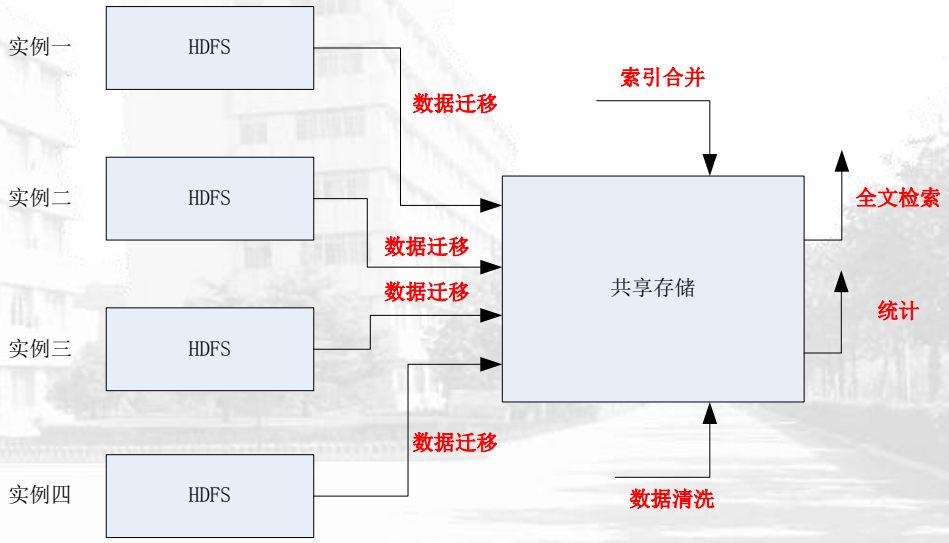
- 数据接收
  - ✓ 支持文件和网络数据流两种方式
  - ✓ 采用数据网关代理模式，将接收到的数据直接保存到HDFS中
  - ✓ 4台接收机，每台机器支持10个进程进行接收和写入，复制因子为3
  - ✓ 经过测试，单机写入速度总带宽为70MB/s

二、系统组成

- 分布式运行支撑环境
  - ✓ 基于hadoop的Map/Reduce
  - ✓ 数据块大小设置为128M
  - ✓ 使用4个hadoop实例，每个实例在每台机器上启动2个任务进程
  - ✓ 经过测试，单机处理速度达到4MB/s

二、系统组成

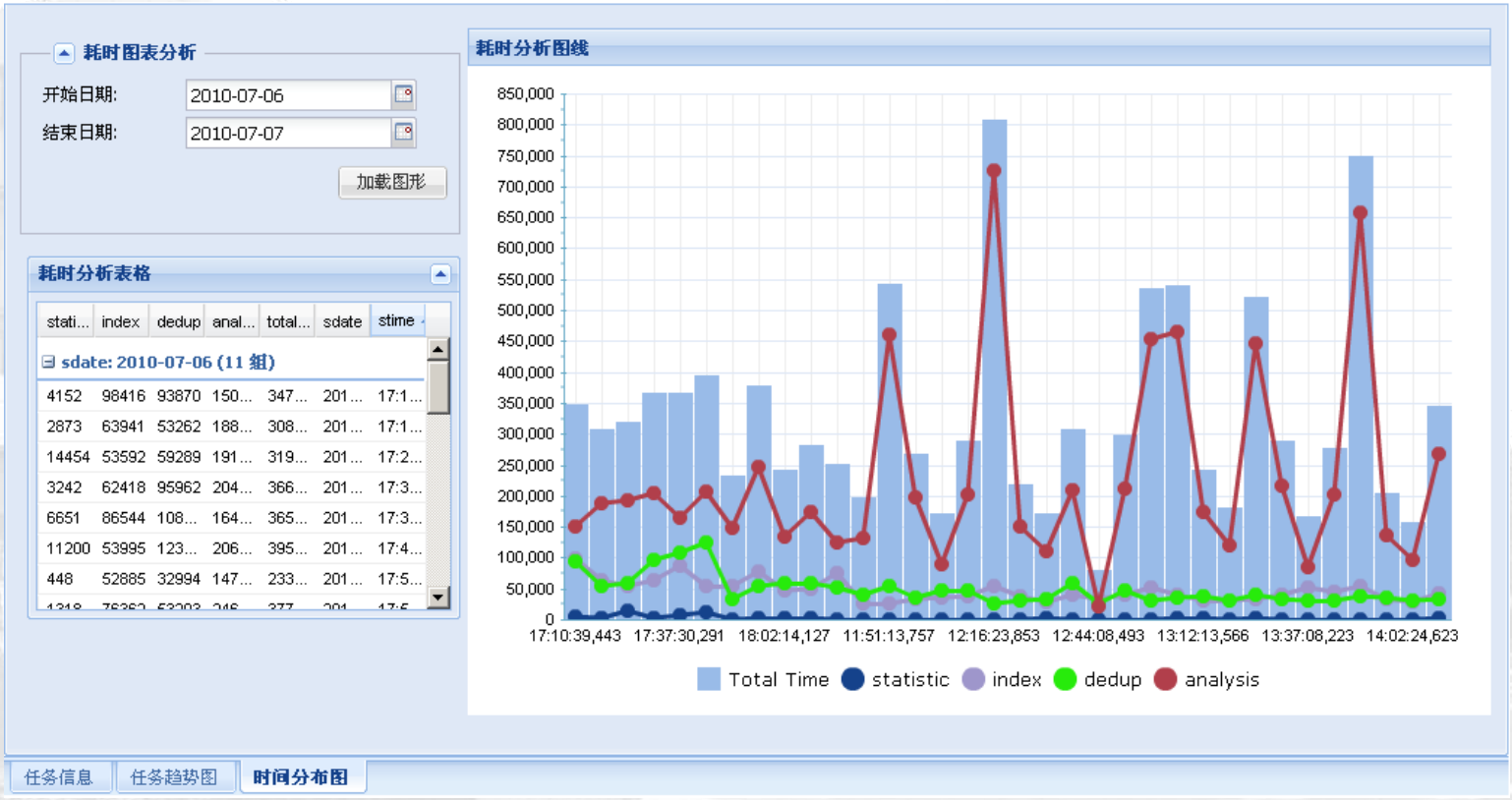
- 应用服务
  - ✓ 数据迁移服务
  - ✓ 索引合并服务
  - ✓ 数据清洗服务
  - ✓ 全文检索服务
  - ✓ 统计服务



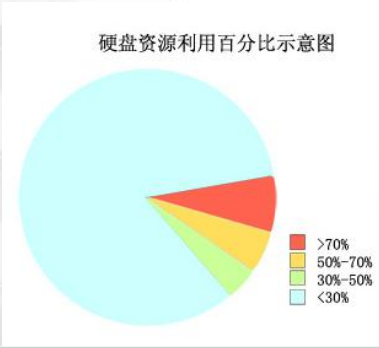
三、系统演示


数据接收
用户:dreceiver
节点:cn125-cn128
cn125
运行中...
cn126
运行中...
cn127
运行中...
cn128
运行中...
基础环境
主结点:cn122
运算结点:cn123-cn124
用户:dguest1 dguest2 dguest3 dguest4
运行中... dguest1:9060
运行中... dguest2:9070
运行中... dguest3:9080
运行中... dguest4:9090
运行管理
主结点:cn122
运算结点:cn123-cn124
用户:dguest1 dguest2 dguest3 dguest4
停止 dguest1:9060
运行中... dguest2:9070
运行中... dguest3:9080
运行中... dguest4:9090
网络服务
查询服务 cn121
运行中...
iipg:8080
推送服务 cn121
运行中...
iipg:8080
WSA 服务
cn121
运行中...
cn121
运行中...

三、系统演示

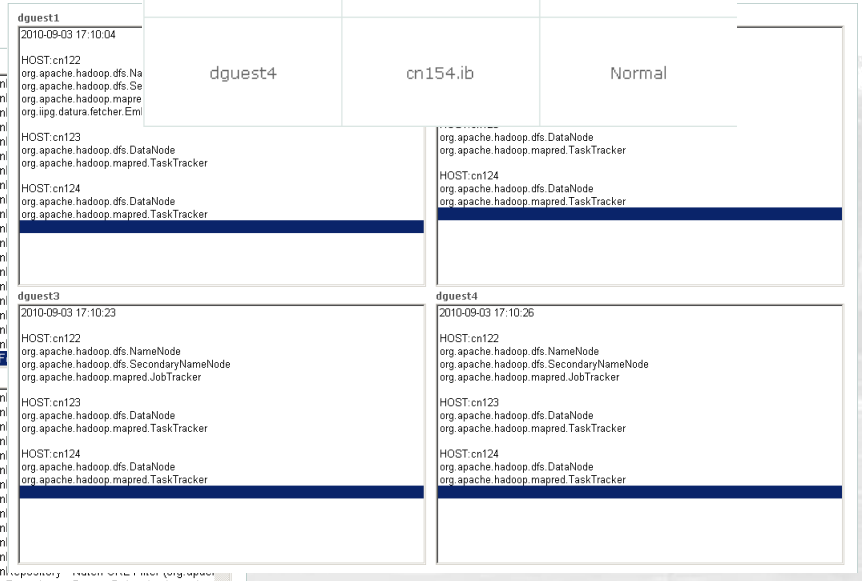

耗时图表分析
开始日期:
2010-07-06
结束日期:
2010-07-07
加载图形
耗时分析表格
stati... index dedup anal... total... sdate stime
sdate: 2010-07-06 (11 组)
4152 98416 93870 150... 347... 201... 17:1...
2873 63941 53262 188... 308... 201... 17:1...
14454 53592 59289 191... 319... 201... 17:2...
3242 62418 95962 204... 366... 201... 17:3...
6651 86544 108... 164... 365... 201... 17:3...
11200 53995 123... 206... 395... 201... 17:4...
448 52885 32994 147... 233... 201... 17:5...
耗时分析图线
850,000
800,000
750,000
700,000
650,000
600,000
550,000
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
17:10:39,443
17:37:30,291
18:02:14,127
11:51:13,757
12:16:23,853
12:44:08,493
13:12:13,566
13:37:08,223
14:02:24,623
Total Time
statistic
index
dedup
analysis
任务信息
任务趋势图
时间分布图

三、系统演示



cn124	cn123	cn122
11%	11%	11%

用户	结点	状态
dguest1	cn154.ib	Unormal
dguest2	cn154.ib	Normal
dguest3	cn154.ib	Normal
dguest4	cn154.ib	Normal



四、前期工作小结

- 针对0.12版本的Hadoop进行的高可用性修正
- 从Nutch的Framework出发创建了文本流数据处理的基础框架
- 开发了Eclipse插件来简化应用开发和部署

五、下一步工作

- 基于Hadoop的优化，包括：
  - ✓ 自动化的数据均衡
  - ✓ 多道任务间的高效数据交换
  - ✓ 实现任务流水化调度机制
- 面向云服务器使用模式开发Eclipse插件，支持应用的开发和调试


ANY QUESTION?