

基于Hadoop的 海量文本处理系统介绍

江南计算技术研究所

江南计算技术研究所
JIANGNAN
INSTITUTE OF COMPUTING
TECHNOLOGY

工作重地
闲人免进

纲要



系统简介



系统组成



系统演示



前期工作小结



下一步工作

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

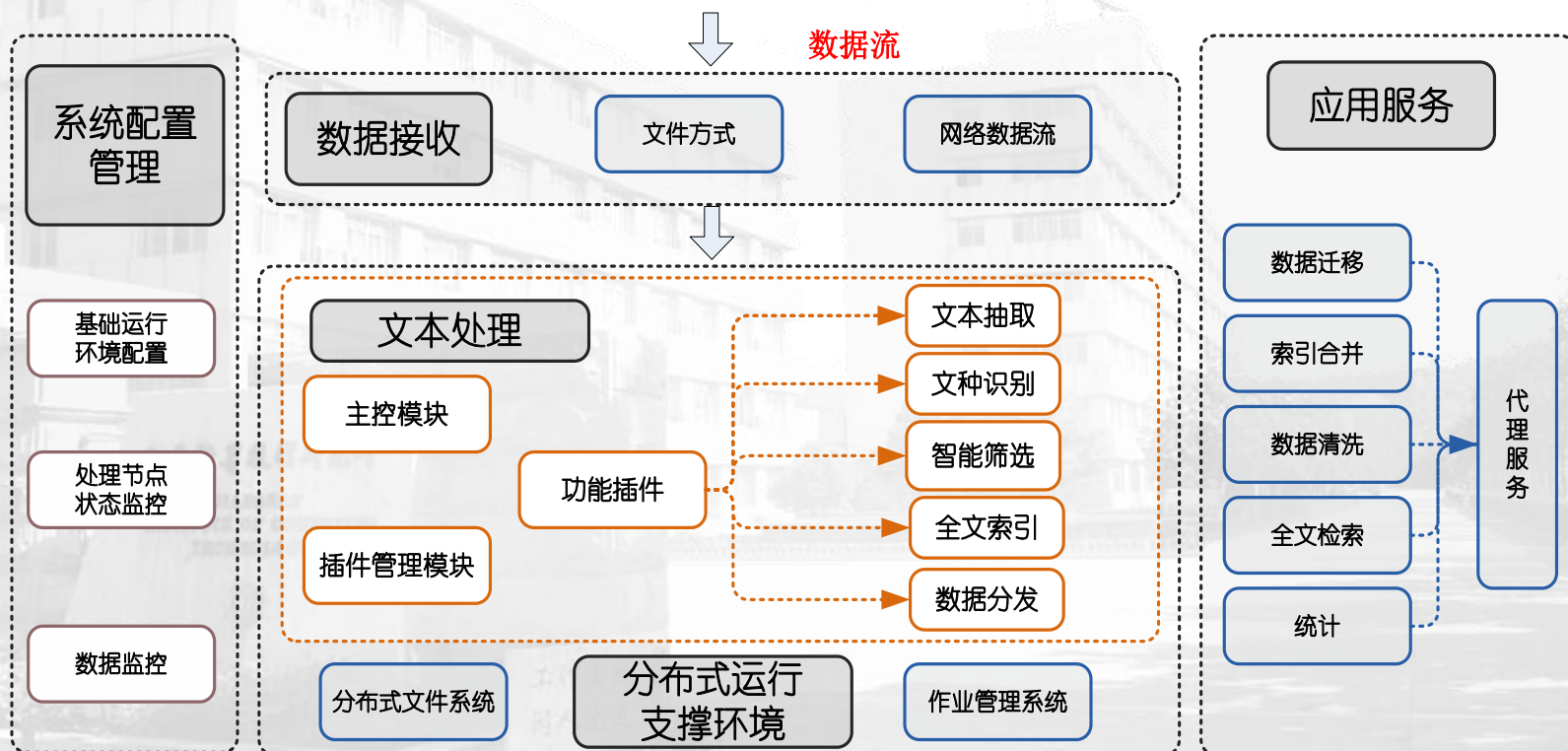
- ✓ Java(TM) SE Runtime Environment (build 1.6.0_13-b03)
- ✓ gcc version 3.4.6 20060404 (Red Hat 3.4.6-10)

➤ 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



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

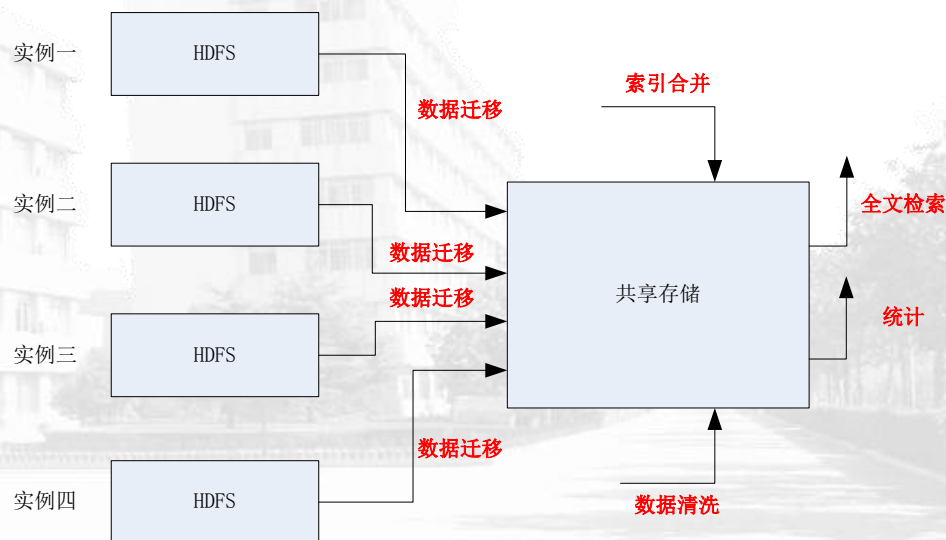
➤ 分布式运行支撑环境

- ✓ 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



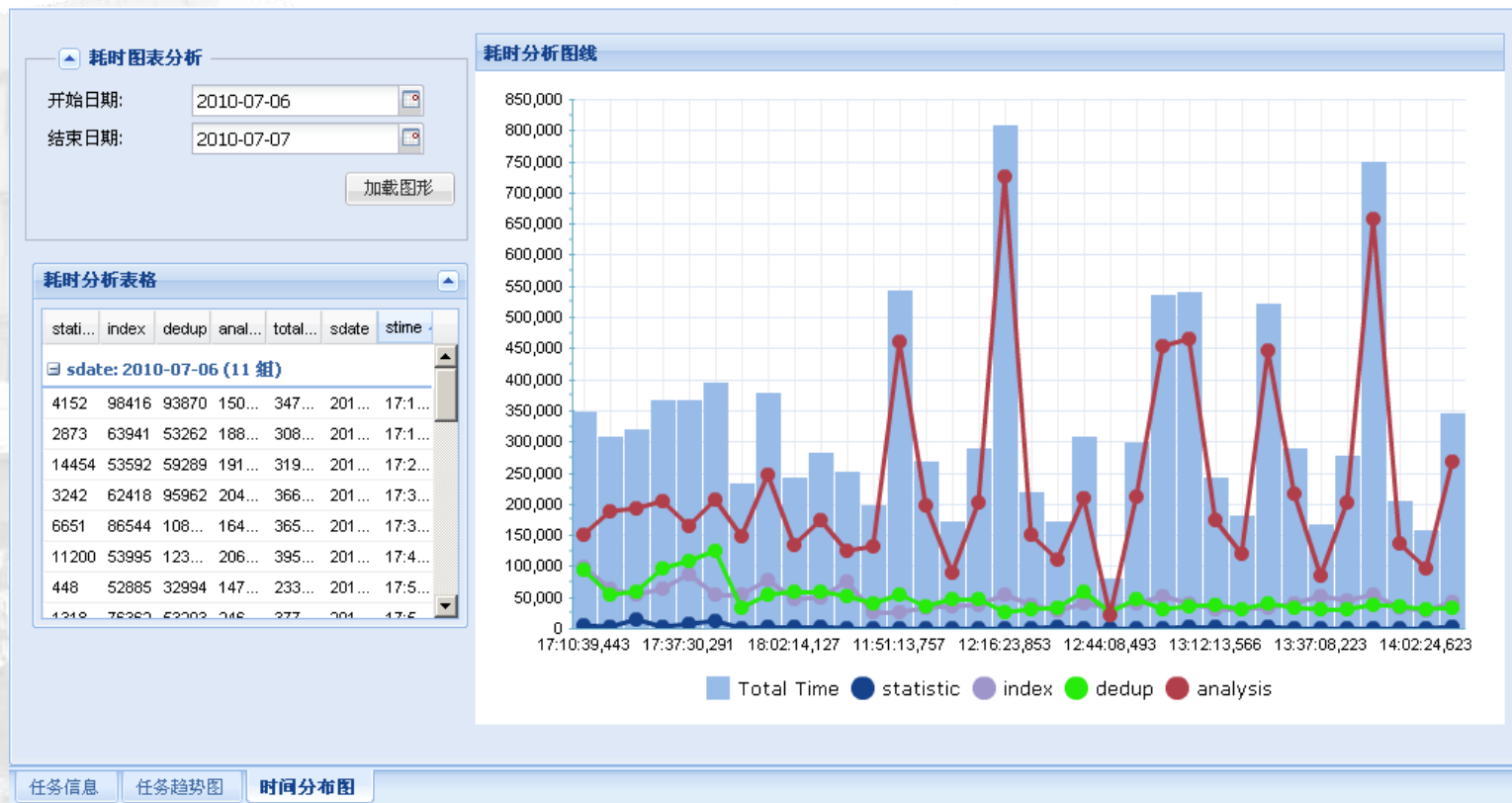
运行中...

cn121



运行中...

三、系统演示





```

dquest1
2010-09-03 17:10:04
HOST:cn122
org.apache.hadoop.dfs.N
org.apache.hadoop.dfs.S
org.apache.hadoop.mapr
org.iip.datura.fetcher.En
HOST:cn123
org.apache.hadoop.dfs.D
org.apache.hadoop.mapr
HOST:cn124
org.apache.hadoop.dfs.D
org.apache.hadoop.mapr

```

```
dquest3
2010-09-03 17:10:23
HOST:cn122
org.apache.hadoop.dfs.N
org.apache.hadoop.dfs.S
org.apache.hadoop.mapr
HOST:cn123
org.apache.hadoop.dfs.D
org.apache.hadoop.mapr
HOST:cn124
org.apache.hadoop.dfs.D
org.apache.hadoop.mapr
```

DataNode red.TaskTracker	<pre> org.apache.hadoop.dfs.DataNode org.apache.hadoop.mapred.TaskTracker HOST:cn124 org.apache.hadoop.dfs.DataNode org.apache.hadoop.mapred.TaskTracker </pre>
NameNode SecondaryNameNode red.JobTracker	<pre> 2010-09-03 17:10:26 HOST:cn122 org.apache.hadoop.dfs.NameNode org.apache.hadoop.dfs.SecondaryNameNode org.apache.hadoop.mapred.JobTracker HOST:cn123 org.apache.hadoop.dfs.DataNode org.apache.hadoop.mapred.TaskTracker HOST:cn124 org.apache.hadoop.dfs.DataNode org.apache.hadoop.mapred.TaskTracker </pre>

- [Model Search Result](#)
- [Model Search Filter \(org.apache.hadoop.hive.ql.exec.tez\)](#)
- [Model Parser \(org.apache.hadoop.hive.ql.exec.tez\)](#)
- [Modeling \(org.apache.hadoop.hive.ql.exec.tez\)](#)
- [Model Loader \(org.apache.hadoop.hive.ql.exec.tez\)](#)
- [Model for another...](#)

dquest3			
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Java attachencrypt connect
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Registered ExtensionPoint
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch Optimizer (org.apa
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch Protocol (org.apa
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch Analysis (org.apa
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Datura Scoring (org.ipg
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Datura Special Filter (org.i
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Datura Dispatch Filter (org
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - HTML Parse Filter (org.ap
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch Query Filter (org.ap
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch URL Normalizer (org
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Datura Email Indexing Fil
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch URL Filter (org.ap
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Datura String Analyzer (org
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch Inline Search Result
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch Content Parser (org
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Nutch Scoring (org.apa
2010-09-03	17:08:20.425	INFO	plugin.PluginRepository - Ontology Model Loader (org
2010-09-03	17:08:21.316	INFO	fetcher.EmailFetcher - No more data, wait for another

dugtest4		org.apache.hadoop.mapreduce.lib.input.FileInputFormat\$FileSplit org.apache.hadoop.mapred.JobTracker
2010-09-03 17:08:18.492 INFO plugin.PluginI	HOST:cn123	
2010-09-03 17:08:18.492 INFO plugin.PluginI	org.apache.hadoop.dfs.DataNode	
2010-09-03 17:08:18.492 INFO plugin.PluginI	org.apache.hadoop.mapred.TaskTracker	
2010-09-03 17:08:18.492 INFO plugin.PluginI	HOST:cn124	
2010-09-03 17:08:18.492 INFO plugin.PluginI	org.apache.hadoop.dfs.DataNode	
2010-09-03 17:08:18.492 INFO plugin.PluginI	org.apache.hadoop.mapred.TaskTracker	
2010-09-03 17:08:18.492 INFO plugin.PluginI		
2010-09-03 17:08:18.492 INFO plugin.PluginI		
2010-09-03 17:08:18.492 INFO plugin.PluginI		
2010-09-03 17:08:18.492 INFO plugin.PluginI		
2010-09-03 17:08:18.492 INFO plugin.PluginI		
2010-09-03 17:08:18.492 INFO plugin.PluginI		
2010-09-03 17:08:18.492 INFO plugin.PluginI		
2010-09-03 17:08:18.492 INFO plugin.PluginRepository - DataString Analyzer (org.apac		
2010-09-03 17:08:18.492 INFO plugin.PluginRepository - Nutch Online Search Result		
2010-09-03 17:08:18.492 INFO plugin.PluginRepository - Nutch Indexing Filter (org.apac		
2010-09-03 17:08:18.492 INFO plugin.PluginRepository - Nutch Content Parser (org.apac		
2010-09-03 17:08:18.492 INFO plugin.PluginRepository - Nutch Scoring (org.apache.		
2010-09-03 17:08:18.492 INFO plugin.PluginRepository - Ontology Model Loader (org.apac		
2010-09-03 17:06:19.960 INFO fetcher.ErmfFetcher - No more data, wait for another		

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?

江南大学
JIANGNAN
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

工作繁忙
闲人免进