XXXX招投标大数据文档

取解析demo说明

目录

[1. 文档说明 1](#_Toc411880964)

[2. 文档对象 1](#_Toc411880965)

[3. 术语定义 1](#_Toc411880966)

[3.1. 技术术语 1](#_Toc411880967)

[3.2. 大数据术语 1](#_Toc411880968)

[4. Demo功能说明 2](#_Toc411880969)

[4.1. Demo登录 2](#_Toc411880970)

[4.2. Demo项目上传 2](#_Toc411880971)

[4.3. Demo项目查询 3](#_Toc411880972)

[4.4. Demo项目管理 4](#_Toc411880973)

[5. 总体技术架构 6](#_Toc411880974)

[5.1. 重要组件描述——结构数据存储 7](#_Toc411880975)

[5.2. 重要组件描述——文档解析 7](#_Toc411880976)

[5.3. 重要组件描述——数据查询 8](#_Toc411880977)

[5.4. 重要组件描述——数据冗余 8](#_Toc411880978)

[6. 数据流程图 9](#_Toc411880979)

[7. 数据存储视图 10](#_Toc411880980)

[7.1. HDFS分布式文件系统 10](#_Toc411880981)

[7.2. HBase NoSQL数据库 11](#_Toc411880982)

[7.3. SolrCloud 集群索引 11](#_Toc411880983)

Pig学习经验

1. Pig概要
2. Pig基本语法
3. 常见在A=load之间必须有分隔符。例如：A = load
4. Pig store(“[local|hdfs：目录而不是文件]”),但是所存储的文件夹必须不存在。输出是以文件目录格式的形式；它会先验证文件是否存在；
5. PigStorage用在文件加载与文件存储中，其中需要注意的是：1. PigStorage默认是以‘\t’作为Field分隔符。2.PigStorage的Tuple,Bag,Map采用内部的分隔符采用的是’,’(逗号)。3. 数据格式：Tuple（Field[,Field…]）,

Bag(Tuple[,Tuple…]),Map[Key#Value[,Key#value….]。4. 定义格式：Tuple=> t\_tuple:tuple(a:int,b:datetime)

Bag=> t\_bag:bag{t:(a:int, b:int)} // 注意在定义的时候一个Bag只能包含一个一层的Tuple。可以包含任意的第二层Tuple。调用时：t\_bag.a 不是 t\_bag.t.a; Map=>t\_map:map[int] ;//数据类型；

Pig自有函数

Pig用户自定义函数

Pig脚本嵌套

Pig性能与效率

Pig测试

数据存储视图

* 1. 图 71.-1 数据存储视图

如上图所示，本项目存储主要分为三层，处于最底层的是源数据层，源数据层主要包括结构化、非结构化数据，这些数据来自于项目输入数据。

第二层是数据WEB转发层，它是用来解决集群的安全性和实际情况而设计的。从源数据层获得数据并进行缓冲。

第三层是数据存储层，它用于存储从平台获取的各类包括结构化、半结构化、非结构化的数据。在数据存储层中，包含5种数据库用于满足不同场景的需求，下面分别阐述其使用场景：

* 1. HDFS分布式文件系统

Hadoop提供分布式的文件系统擅长存储结构化、半结构化、非结构化的数据。在数据存储时，由于HDFS采用的是Share-nothing架构，多节点可以同时进行数据的存储，具有极高的数据吞吐量，数据存储和读取速度得到大大提升。在本系统中，hdfs中存储的内容包括：

1. 文件或是不需要频繁更新的大块数据（例如单文件1M~10G），最典型的是Word、Excel等招投标文件，存储在hdfs分布式文件系统中，其hdfs的访问路径存储在关系型数据库或HBase中。
2. 项目文件的历史数据通过压缩的方式进行存储。
   1. HBase NoSQL数据库

HBase是基于列进行存储的NoSQL数据库，支持列的动态添加，HBase能自动切分数据，并将数据存储在Hdfs的DataNode节点上，使得HBase的存储具有自动具有水平扩展的能力。HBase通过Master节点，可以循序定位到数据存储所在Region节点，并能够支持多线程读取，具有高并发读写能力。能够通过不停机动态扩展Region节点，获取更高的并发处理性能。HBase可以在百亿级别的记录中，在毫秒级别时间内查询出指定ROWKEY的数据。

在本次系统建议方案中，HBase存储文档的元数据以及文档在hdfs上存储的路径，用于快速检索并提取文档。同时HBase用于存储业务数据，例如招投标信息和交易记录，而这些数据的二级索引，通过SolrCloud进行支持。

* 1. SolrCloud 集群索引

SolrCloud是系统核心检索组件，提供全文检索能力，在系统中，文档的元数据在HBase中存储，并在SolrCloud中建立索引，同时，文档被解析后，形成文本摘要，通过SolrCloud形成关键字索引，提供给用户按照关键字检索查询的能力。

需要检索的业务数据，也可以在Solr中进行索引以提供高性能检索的能力。