# 大数据平台服务手册

1. 服务器节点服务分布

Hadoop节点：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Hostname | Namenode | Secondnamenode | Datanode | Resourcemanager | Nodemanager |
| slave1 | √ | √ | √ | √ | √ |
| slave2 |  |  | √ |  | √ |

HBase节点：

|  |  |  |
| --- | --- | --- |
| Hostname | HBaseMaster | HBaseRegionServer |
| slave1 |  | √ |
| slave2 | √ | √ |

Spark节点：

|  |  |  |  |
| --- | --- | --- | --- |
| Hostname | SparkMaster | SparkWorker | SparkHistoryServer |
| slave1 | √ | √ | √ |
| slave2 |  | √ |  |

Kafka节点：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hostname | KafkaBroker | Kafka Manager | KafkaProducer | KafkaConsumer |
| slave1 |  | √ |  |  |
| slave2 | √ |  |  | √ |

Zookeeper节点：

|  |  |
| --- | --- |
| Hostname | url |
| slave1 | 192.168.18.142 |
| slave2 | 192.168.18.143 |

1. 服务访问地址

Hadoop节点管理地址：192.168.18.142:8088

HBase管理地址:192.168.18.143:16010

HBase查询地址:192.168.18.143:2181

Spark应用管理地址：192.168.18.142:8088/cluster/apps

Spark历史服务器地址:192.168.18.142:18080

Kafka管理地址:192.168.18.142:9000

1. 服务简介

Hadoop节点管理：

提供hadoop节点的生存状态，包括运行状态、内存分配、核心使用、应用管理以及日志管理。

HBase管理:

提供hbase节点状态，包括table表信息、region信息、任务请求信息以及日志管理；提供table表的操作，包括合并和分割region。

HBase查询地址：

使用HBase客户端API进行查询的接口。

Spark应用管理：

提供正在运行的spark应用的状态，包括核心运行的task、task运行的job、job的DAG图，Streaming应用的输入速率、调度延迟以及运行时间。

Spark历史服务器：

提供已经运行完成、失败的应用历史信息。

Kafka管理:

提供kafka节点的运行状态，包括节点的topic、producer、consumer的信息；提供kafka的操作，包括创建topic、修改配置等。

1. HBase表设计

HBase表以rowkey主键、columnFamily列簇和qulifier列组成。

Rowkey作为索引，对查询性能影响最大，如果查询时可以选定rowkey的范围可以大大提升查询的速度；rowkey的长度对查询影响较大，越短越好，一般为16的倍数以对齐内存。

ColumnFamily保存相关联的列，一个表中的ColumnFamily越少越好，命名的长度越短越好。

Qualifier以Byte保存列信息，对于数据的格式由开发人员管理；列为动态生成，表中每一行数据可以由不同列构成，简而言之，HBase表为稀疏表。

test2表设计：

<>代表ptd\_dataset中的字段

|  |  |  |
| --- | --- | --- |
| Rowkey | ColumnFamily | qulifier |
| [buchet]<ts.end><id> | v | 首字段名，如<ts>,<url> |

对于嵌套的字段，如<ts.end>,暂时直接保存在<ts>,查询时需要注意。

存入的数据为原始json字符串，如{"end":"2018-06-25 12:29:55","start":"2018-06-25 12:29:55"}。  
 <ts.end>中的符号如”:”,”-“和空格将去掉。

[buchet]为1byte大小的分片标志，由<ts.end><id> % buchet\_size得到。设置buchet可以使数据均匀分布到各个HBase节点中。

对于长度不到32的rowkey，需要填充字符。

1. HBase表查询

使用scan获取多条数据，get获取单条数据。

scan.withStartRow和scan.withStopRow设置rowkey范围。

scan.setFilter来设置过滤器，过滤不需要的数据，类似SQL的条件查询

具体可以使用的过滤器可以查看<https://hbase.apache.org/book.html#client.filter>

//TODO

编写可供外部简单调用的接口