Hadoop MapReduce手机流量统计

分别使用hadoop的Mapreduce和hive的功能来统计

【1、mapreduce】

1. 分析业务需求：用户使用手机上网，存在流量的消耗。流量包括两个部分：其一为上行流量（发送信息流量），其二为下行流量（接受信息流量）。每种流量在网络传输过程中，有两种形式说明：包的大小，流量的大小。使用手机上网，以手机号为唯一标识符，进行记录。有记录，包括很多信息。



如下所示：



然而我们实际需要的字段：

手机号码、上行数据包数、下行数据包数、上行总流量、下行总流量

【我们Mapreduce的数据已经经过了清洗，不是原数据。】

1. 自定义数据类型（五个字段）（之所以不用数据段拼接，是因为它耗内存）

DataWritable 实现 WritableComparable接口。

（3）分析MapReduce写法，哪些业务逻辑在Map阶段执行，哪些业务逻辑在Reduce阶段执行。

Map阶段：从文件中获取数据，抽取出需要的五个字段，输出的Key为手机号码，输出的Value为数据量的

类型阶段DataWritable对象。

Reduce阶段：将相同手机号码的Value中的数据流量进行相加，得出手机流量的总数（数据包和数据流量）。

输出到文件中，以制表符分开。

load data inpath /opt/wc/input/HTTP\_20130313143750.dat into table phonemessage;

/opt/wc/input/HTTP\_20130313143750.dat

【附录】

电信的实际业务，产生实际的应用

在实际的业务中，原始数据，存储在文件或者关系型数据库，需要进行多次的数据的清理和筛选，

符合我们的数据需要，将不合格的数据全部进行过滤。

sqoop框架，将关系型数据和HBase、Hive以及HDFS中导入导出数据。

针对复杂的实际业务，往往数据，都要自己手动编写代码和进行数据清洗。

【2、Hive】

hive是hadoop家族成员，是一种解析like sql语句的框架。它封装了常用MapReduce任务，让你像执行sql一样操作存储在HDFS的表。  
hive的表分为两种，内表和外表。  
Hive 创建内部表时，会将数据移动到数据仓库指向的路径；若创建外部表，仅记录数据所在的路径，不对数据的位置做任何改变。

1. 首先根据这张表我们可以知道有十个属性：



因此我们在hive中创建的表phonemessage应该如下：

use database hive;

create table phonemessage(reportTime string,msisdn string,apmac string,acmac string,host string,siteType string,upPackNum bigint,downPackNum bigint,upPayLoad bigint,downPayLoad bigint,httpStatus string)row format delimited fields terminated by '\t';

1. 然后将hdfs目录下的数据文件load到表中，传数据

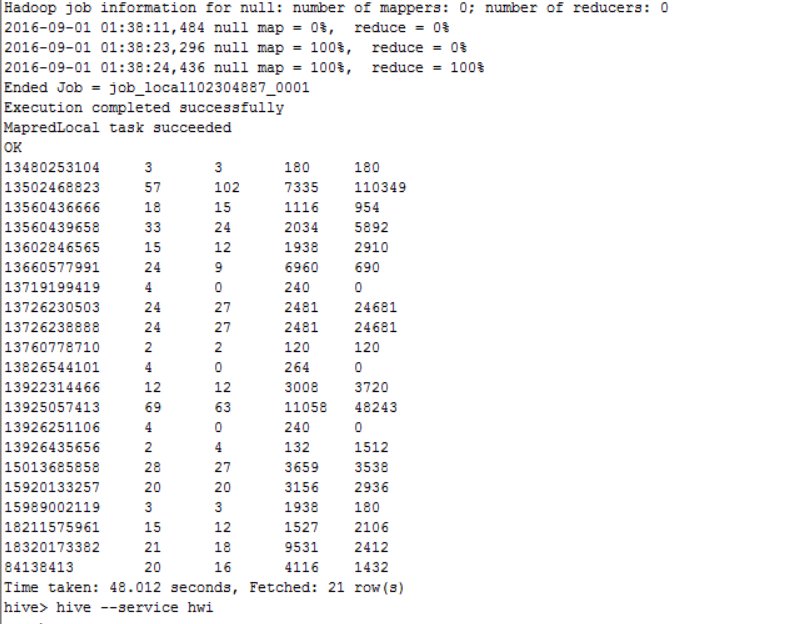
load data inpath '/opt/wc/input/HTTP\_20130313143750.dat' into table phonemessage;

（3）执行hive 的like sql语句,对数据进行统计

Select msisdn,sum(uppacknum),sum(downpacknum),sum(uppayload)

,sum(downpayload) from phonemessage group by msisdn;

得到如下结果：



在hdfs 的/user/hive/warehouse/hive.db/phonemessage/目录下

会有一个HTTP\_20130313143750.dat

Hive的web界面方式下所示如下：

