# 渠道预算模型设计文档

# 模型设计目标

准确预测涉及到3万多个渠道数千个合作商的**每个产品**未来12个月每个月**渠道总预算**量和**独立渠道**预算量。

# 模型设计思路

## 渠道的主要类别

独立渠道，是指在过去一年内某月进入该产品的月渠道（仅限**收费类型**为CPA、CPT或CPM的收费渠道）实际预算量排名前50的渠道。

其他渠道，是指除独立渠道以外的其他渠道，模型不单独测算其预算量。

## 独立渠道的建模方法

为了预测该产品未来十二个月独立渠道的每月预算，需要针对不同类型渠道的不同表现分别建模，具体的类型划分为：

1. 持续上升型
2. 持续下降型
3. 平稳波动型

这三型用MA模型建模

1. 峰值衰减型

这一型用对数函数或有理函数建模

1. 周期波动型

这一型用离散周期函数建模

1. 突然爆发型

这一型用常数函数建模

## 产品预算量的建模方法

产品预算量，主要是按照下面的方式计算：

某产品某月的预计预算量

= 预计该产品本月按照预计预算量排前30的独立渠道的预算量加总 / 预计前30预算量占比

其中：

预计该产品本月按照预计预算量排前30的独立渠道的预算量加总，需要计算所有独立渠道下月的预计预算量，然后进行排序，取前30加总得到；

预计前30预算量占比，需要按照ARIMA模型计算预计前30预算量占比，再**加上**一个外生冲击变量计算得到。

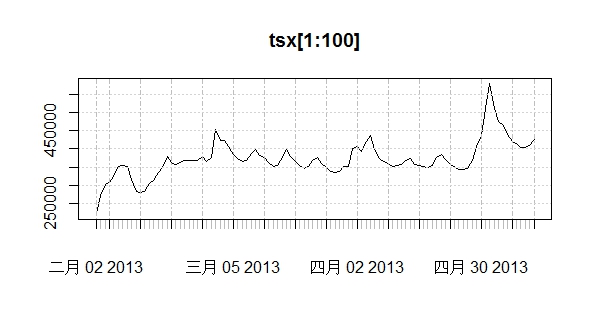
## 按天收费成本数总量的建模方法

历史建模方法中，需要分别对TOP-K 个付费合作方的按天成本数据，以及TOP-K的数据相对于总量数据的百分比（以下简称为：总量占比）进行建模。

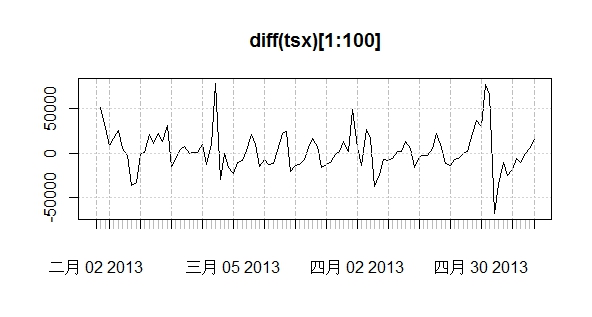
其中，TOP-K数据拟合中，由于成本数据的来源直接，而且业务线涉及到对未来数据的预测，故TOP-K 主要基于自身的历史数据进行回归拟合，优化了拟合的短期效果而不能保证更长时期的效果，而且由于受当前时刻的数据特征的影响而波动很大。对该模型的优化不能有效解决这些问题。

总量占比建模思路采用时间序列方法，但由于TOP-K个付费合作方选择的滞后性等原因导致该总量占比的代表性不够显著，对该模型进行优化对最终产生的总量数据的拟合效果的提升空间受限制。

综上所述，此处重新提出按天收费成本数总量进行建模的方法。通过观测历史按天总量数据的特征，通过对其进行差分变换、 对数变换分析。见下图：

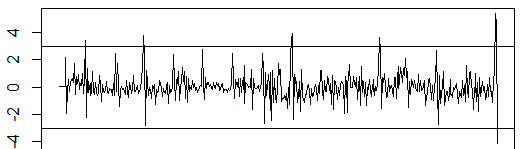


原始数据（取前100天的数据）



对数差分变换后的数据（取前100天的数据）

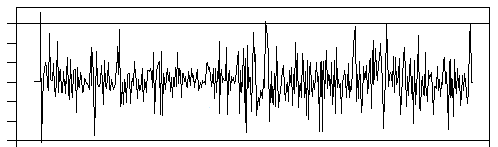
对差分后的数据进一步分析，构造出数据中的周期规律S： ARIMA(p,d,q)(P,D,Q)[S]。通过参考自动识别的阶数，再适当对比多个(p,d,q,P,D,Q)的值，得到最终修正后的模型(4,0,2)(1,1,1)[7]的残差图如下。



残差图

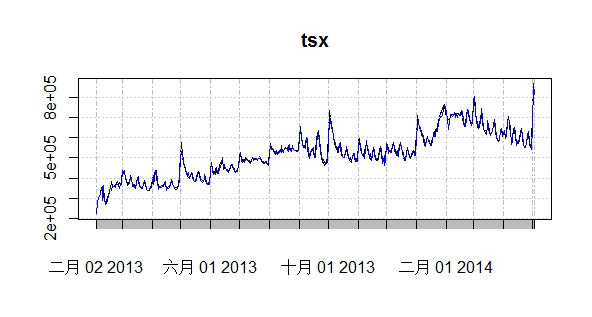
通过分析结果的残差，发现在一些国家的节假日、每月的前几天有明显的大幅度波动。因而，将元旦、五一等节假日、每月的前3天作为特征维度增加到模型中。

继续分析结果的残差，发现在2013年春节前后，五一假期前后有模型无法捕捉的异常。因而，将这些特殊时间点加入到模型中。



修正后的模型的残差图

综上所述，最终的拟合效果图如下：



## 每月计算流程

# 数据模型细节

数据模型的数据表罗列如下：

PRO\_DIM\_DATE

PRO\_DIM\_ORGN

PRO\_INSIGHT\_FACT

这三个表构建之后，初次插入数据后，就一直处于活动状态，均为**每周更新一次**。

这三个表的表结构如下：

**PRO\_DIM\_DATE**

DATE, VARCHAR(10), 日期

YEAR, VARCHAR(4), 哪一年

MONTH, VARCHAR(2), 哪一个月

YEARMONTH, VARCHAR(7), 年月

DAYS, INT, 一年中的第几天

WEEKS, INT, 一年中的第几个星期

WEEKS\_PERCENT, FLOAT, 一年中星期百分比

WEEKORDER, INT, 一周中的星期几

ISHOLIDAY, INT, 是否节假日

HOLIDAYNAME, VARCHAR(20), 节假日名称

**PRO\_DIM\_ORGN**

PRODUCTION,VARCHAR(100), 产品名

COP\_ID,INT, 合作商ID

COP\_NAME,VARCHAR(200), 合作商名称

CHANNEL\_ID,VARCHAR(20), 渠道ID

COP\_TYPE, VARCHAR(20), 合作方类型

COP\_MODE, VARCHAR(20), 合作方模式

COP\_CITY, VARCHAR(20), 合作方地域

INDEX\_NAME,VARCHAR(20), 合同指标名

PRICE, FLOAT, 合作方单价

**PRO\_INSIGHT\_FACT**

DATE,VARCHAR(10), 日期

PRODUCTION,VARCHAR(100), 产品名

COP\_ID,INT, 合作商ID

CHANNEL\_ID,VARCHAR(20), 渠道ID

DATA\_TYPE,VARCHAR(20), 数据类型

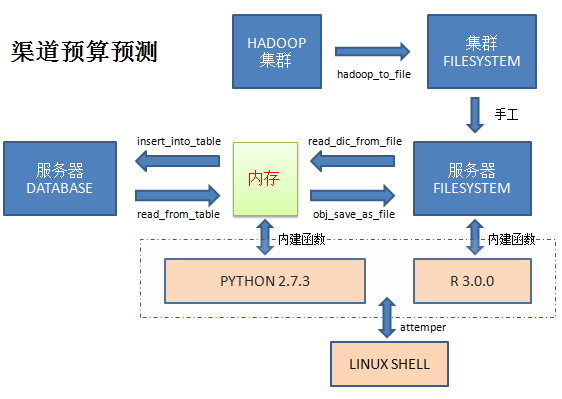
ACTIVE\_NUM,FLOAT, 用户活跃数

ACTIVE\_NUM1,FLOAT, 用户一次激活数

ACTIVE\_NUM2,FLOAT, 用户二次激活数

# 整体计算架构

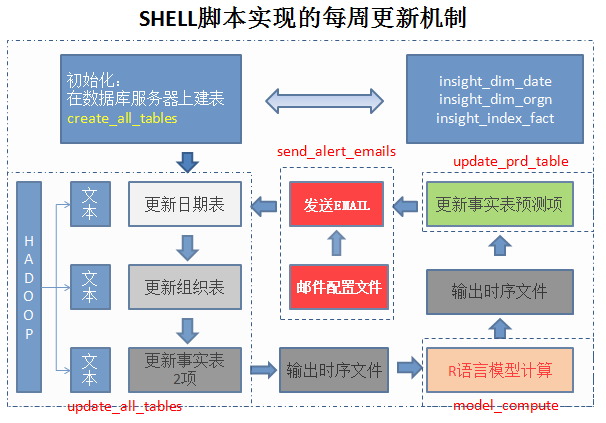
渠道预算模型的整体架构如下：



大部分文件是.py文件，最终SHELL脚本attemper.sh会放入到计划任务中，每周跑一次。

# 数据更新

更新过程的具体步骤是：



每周更新完毕后会发送一封邮件到各相关人员邮箱中。

# 最终的交付物和交付期限

**文档包括：**

项目立项word文档 (2014.01.21)

业务需求word文档 (2014.02.21)

模型设计word文档V1.0 (2014.02.28)

模型设计word文档V2.0 (2014.03.07)

模型设计word文档V3.0 (2014.03.14)

首次预测数据txt样例(2014.03.21)

模型部署word文档 (2014.03.28)

模型维护word文档 (2014.04.04)

**代码包括：**

related\_func.py (2014.03.14)

create\_all\_tables.py (2014.03.14)

update\_all\_tables.py (2014.03.21)

model\_compute.R (2014.03.21)

update\_prd\_table.py (2014.03.21)

send\_alert\_emails.py (2014.03.28)

attemper.sh (2014.03.28)

**初始化数据库文件包括：**

date.txt, orgn.txt, fact.txt (2014.03.28)