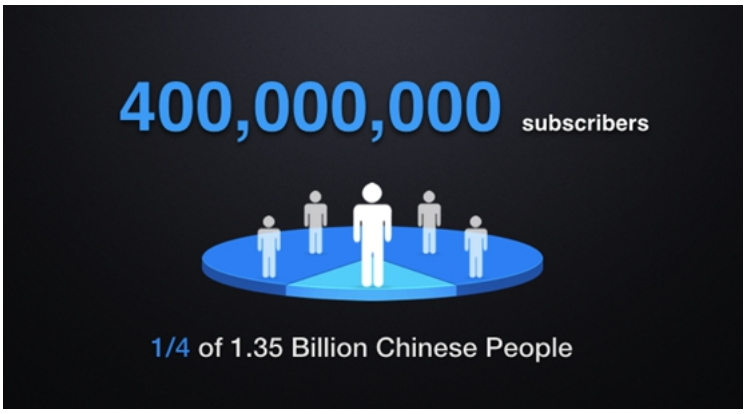
阿里云ODPS实践：墨迹为4亿用户提供个性化天气服务

发表于 2015-12-17 16:17 | 3182次阅读 | 来源 CSDN | 0 条评论 | 作者 章汉龙

大数据 阿里云 ODPS 墨迹天气

摘要：墨迹天气成立到现在5年多，已经积累了4亿用户，4亿是什么概念？13.5亿中国人，每四个人中就有一个下载过墨迹天气，4亿的独立注册用户数超过美国人口总数。

在墨迹天气上，每天有超过 5 亿次的天气查询需求，这个数字甚至要大于 Twitter 每天发帖量。墨迹天气已经集成了多语言版本，可根据手机系统语言自动适配，用户覆盖包括中国大陆、港澳台，日韩及东南亚、欧美等全球各地用户。运营团队每天最关心的是这些用户正在如何使用墨迹，在他们操作中透露了哪些个性化需求。



这些数据全部存储在墨迹的API 日志中，对这些数据分析，就变成了运营团队每天的最重要的工作。墨迹天气的API每天产生的日志量大约在400GB左右，分析工具采用了阿里云的大数据计算服务ODPS。

使用ODPS的逻辑流程如下：

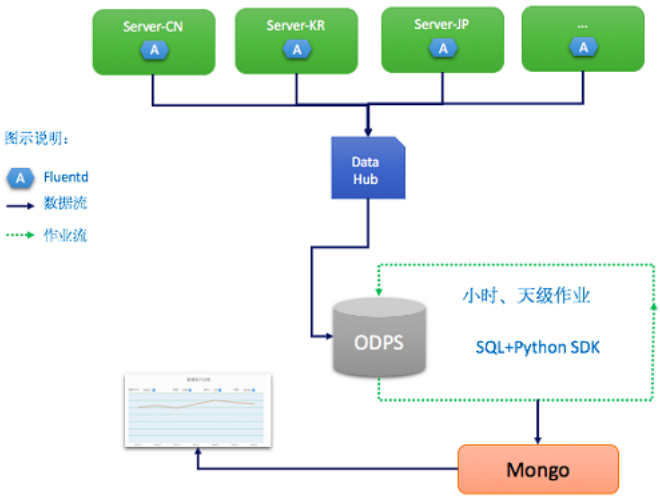


图2 墨迹日志分析流程

流程介绍：

1. 在每个日志服务器上都安装了Fluentd及ODPS数据导入插件。日志数据通过流通道DataHub实时导入到ODPS；

2. 数据分析作业分小时级和天级任务。数据开发工程师通过ODPS Python SDK向ODPS提交SQL 分析脚本，将统计后的数据导入Mongo DB。报表系统直接对接Mongo DB；

3. 运营人员通过报表系统来查看用户统计结果；

整个数据分析过程也做了很多优化。以下是几点说明：

1. 导入工具Fluentd。Fluentd是一款优秀的日志导入软件。代码开源，支持Apache License 2.0。Fluentd支持300多个插件，基本上今天的大数据处理系统，Fluentd都能支持。Fluentd还支持自定义插件，允许通过代码编写其它数据源和目标。使用配置简单、灵活，底层引擎关键部分通过使用C语言类库编写，所以性能比较好。墨迹选择了使用Fluentd向ODPS导入数据。

2. 时区数据的统一。 墨迹的服务器部署在不同时区，日志数据按天和小时两级分区流入到ODPS表中，但统计作业是发生在北京时间。例如，对于2015年12月1日的数据统计是在12月2日凌晨来做的。由于时区不同，统计作业运行完毕后，仍有部分时区在12月1日的数据会持续流入1日的分区表中，这就会导致这部分数据在统计时落掉。

解决这个问题，在实施时将所有的日志数据中的local时间按北京时间做了转换，截止到北京时间12月1日结束时，所有数据流入1日的分区中。其它时区是1日的数据会流入2日的分区，数据会在第二天完成统计。Fluentd中Filter 插件可以完成这个转换操作，配置非常简单，如下面部分代码：

```
<filter filter-tag>
    type record_transformer
    enable_ruby
    <record>
Bjdatetime ${(Time.strptime(LocalDatetime,'%m/%d-
%H:%M:%S,%L').gmtime+8*3600).strftime('%Y-%m-%d %H:%M:%S')}
    </record>
</filter>
```

3. 任务的调度。墨迹分析的作业每天和每小时都会执行。分析后的数据导入本地Mongo DB，报表系统接入Mongo DB来做展现。墨迹分析工程师在本地使用定时调度Python脚本完成这些流程。SQL 分析脚本可以通过ODPS Python SDK直接提交到ODPS上执行完，完成后将统计结果放到List 对象。通过Python Mongo Client 将List写入Mongo DB。

墨迹天气的这一流程之前是在国外某云计算平台上完成的，需要分别使用云存储、大数据分析等服务，数据分析完成后再同步到本地Mongo DB中与报表系统对接。在迁移到ODPS后，流程上做了优化，EMR的工作省掉了，日志数据导入到ODPS表后，通过SQL进行分析，完成后直接将结果写入本地Mongo DB。

在存储方面，ODPS中的表按列压缩存储，更节省存储空间，整体上存储和计算的费用比之前省了70%，性能和稳定性也提高了很多。同时墨迹可以借助ODPS上的机器学习算法，对数据进行深度挖掘，为用户提供个性化的天气服务。

作者简介：章汉龙　墨迹天气运维部经理

推荐阅读相关主题： 云计算平台　数据分析　机器学习　手机系统　工程师　服务器

相关文章　最新报道

贵州大数据共享平台“云上贵州”2000万重奖征集大数…
天脉聚源：如何用1200台阿里云支撑春晚“倒计时”
优化无极限：盘古Master优化实践
阿里云课堂第六期：大型互联网应用架构之存储与分发
访阿里云大规模存储“铁三角”：OSS、RDS与OTS
走近华佗，解析自动化故障处理系统背后的秘密