# 绪 论

## 课题背景及选题意义

随着4G、Wi-Fi等移动通信技术的飞速发展，移动互联网行业的时代已经到来。作为用户进入互联网的重要入口之一，手机应用程序在移动互联网中已经占据了举足轻重的位置。根据iiMedia Research（艾媒咨询）2016年7月29日发布的《2016上半年中国手机桌面行业研究报告》显示截止2016年第二季度，中国智能手机用户已达6.308亿人。庞大的用户基础推动了手机应用程序的快速发展。手机应用程序承载了各种便捷的移动服务，已经成为人们日常生活的一部分。手机应用程序的性能监控与应急处理已经受到移动互联网行业的重视。

为了更好的服务用户，了解用户的访问特征及用户需求，了解手机应用程序的运行情况以及性能情况手机应用厂商迫切需要对用户的使用情况进行分析研究。现在各大手机应用程序公司的普遍做法是记录用户使用情况，记录产品运行情况，然后根据产品运行数据和用户使用行为数据进行分析，进而改进产品，给用户带来更好的体验。应用程序新版本在发布之前的性能测试也已经成为另一个非常重要的工作，通过产品的总体性能测试数据和各个模块分别的性能测试数据，开发人员可以快速发现并定位产品中存在的问题和相关模块的性能瓶颈。另一方面，产品性能的测试数据也是衡量产品是否可以发布的重要指标。

产品日志是手机应用软件在运行过程中对自身运行状态进行实时记录的文件，通常这些文件还包括系统的所有操作、软件自身行为以及用户的行为。这些数据通常用于软件故障排查和软件性能分析，在数据量不多时，日志的处理和分析由经验丰富的产品开发工程师人工查找完成，但是随着系统规模的扩大和用户数量的增加，日志的数量正以前所未有的速度快速增长，这种情况下继续采用人工的方式进行处理将是一件不可能完成的任务。因此，为了高效的处理大规模日志，解决单机系统上的存储和处理瓶颈，提高处理效率，各大公司都在开发研究分布式日志管理系统和相应的软件性能分析系统[1]。

使用软件性能分析系统可以对大量的日志数据进行智能分析，自动计算产品性能，方便诊断和优化产品的缺陷。软件性能分析系统可以量化的衡量整个产品的质量，通过分析产品中各个模块的数据，可以直观得到每个模块的性能数据进而分析出产品的性能短板；通过分析产品不同流程不同模块的数据，可以更快的发现和定位产品的问题，使开发工程师有的放矢；也可以分析用户数据，分析用户使用习惯，分析用户喜好，对用户进行分类，并针对每种不同类型的用户推送其喜欢的话题和功能，这样可以为下一步产品的改进提供建设性的意见。

因此，在微软小娜产品发布之初，我们开发了微软小娜性能分析系统。目的是采用大数据处理的方法更高效更便捷的分析产品日志，为微软小娜产品的发布和新功能的开发提供可以量化的性能指标，并且将这些指标实时的以可视化的方法展示给开发人员。

## 国内外技术现状

随着互联网的迅猛发展，如今的互联网企业每天需要处理数以万计甚至百万千万的用户访问请求。随着系统规模的扩大和用户数量的增加，用户的日志数量更是成几何倍数增加，这种高并发大数据的日志发送接收和处理需要用到分布式软件系统(Distributed Software Systems) [2]。分布式软件系统是在由通信网络互联的多处理机体系结构上支持分布式处理任务的软件系统。分布式软件系统包括分布式操作系统、分布式程序设计语言及其编译系统、分布式文件系统和分布式数据库系统等。在日志收集中主要使用的是分布式文件系统，分布式文件系统主要解决了数据的存储和管理问题：将原本固定在某一个地点的某一个文件系统扩展到任意多个地点或多个文件系统，将这些节点组成一个方便统一管理的文件系统网络。每个节点可以分布在不同的地点，节点间通过网络进行通信和数据传输[3]。在使用分布式文件系统时，不需要关心数据是存储在哪个节点哪个地点上、或者是从哪个节点那个地点获取的，只需要和使用本地文件系统一样管理和存储文件系统中的数据。

为了方便快速收集处理大规模并发数据，分布式日志收集系统已经得到了广泛应用。其中，Scribe是Facebook公司的开源日志收集处理系统，主要采用“分布式收集，统一处理”的可扩展高容错的方案，从各个日志源上收集日志并将日志存储到分布式文件系统上然后再统一处理。Chukwa是Apache公司的一个开源的大型分布式系统主要用于日志的收集和分析[4]。Chukwa是构建在Hadoop的HDFS和MapReduce框架之上的数据采集分析框架，其继承了Hadoop的可伸缩性和鲁棒性[5]。该框架提供采集数据的客户端（Client），由客户端发送数据给服务器（Server），然后通过定期运行脚本分析数据。Flume是由Cloudera公司提供的一个分布式可靠的海量日志采集、聚合和传输系统[6]。Flume支持在日志系统中定制各类数据发送方用于收集数据，同时对数据进行简单处理并且可以输出到各种数据接收方。Aria是微软（Microsoft）公司开发的Cosmos之上的日志收集系统，作为共享数据平台的一部分，Aria可以支持平均每秒5百万次日志收集。将日志导入Cosmos之后，采用Scope脚本更加详细的计算所需数据。Aria支持实时多维的大数据分析，用户可以创建展示页面，展示多维立体数据并且可以实时查看原始日志。

在大型软件系统中，由于系统环境设置和软件规模等因素，错误（bug）重现和调试都十分困难。百万用户级别的访问量每天会产生上百乃至上千兆字节大小的日志，为了能够高效快速的处理大量日志分布式大数据处理平台广泛应用于日志处理系统。微软主要采用Scope脚本语言对大规模用户日志进行处理。Scope语言是运行在Cosmos上的查询语言。Cosmos是微软研发的大数据处理平台。Cosmos主要有以下两大基本功能：（1）使用分布式文件系统CFS(Cosmos File System)存储大量数据的功能；（2）使用Scope语言处理大数据的功能。Cosmos的只要特点如下：（1）大规模数据处理能力，Cosmos支持PB级别的数据处理和大量可信赖的计算机集群；（2）Cosmos可以使处理大数据更加容易，Scope语言简单易学，允许我们在不懂得MapReduce细节的情况下，使用简单SQL和C#知识就可以进行大数据的运算处理；（3）Cosmos也是一个共享数据系统 (Shared Data Ecosystem)，在工作中不同的组可以在系统提供的平台上相互共享数据。

SQL Server是微软推出的关系型数据库，是一个可扩展的、高性能的、为分布式客户机/服务器计算所设计的数据库管理系统，实现了与Windows NT的有机结合，提供了基于事务的企业级信息管理系统方案。它最初是由Microsoft、Sybase和Ashton-Tate三家公司共同开发的，并于1988年推出了第一个OS/2版本。Microsoft SQL Server近年来不断更新版本，1996年，Microsoft 推出了SQL Server 6.5版本；1998年，SQL Server 7.0版本和用户见面；SQL Server 2000是Microsoft公司于2000年推出，目前最新版本是2012年3月份推出的SQL SERVER 2012[7]。

本系统网站是采用IIS服务部署的，IIS（Internet Information Services，互联网信息服务）是由微软公司提供的基于Microsoft Windows操作系统的互联网基本服务。IIS是一种网页服务组件，其中包括FTP服务器、NNTP服务器、HTTP/HTTPS服务器、Web服务器和SMTP服务器，分别用于文件传输、新闻服务、超文本传输服务、网页浏览和邮件发送服务。

## 本文研究内容

小娜性能分析系统包括从用户操作产品产生日志开始，经过Aria收集日志并将日志导入到Cosmos分布式文件系统中，再使用Scope语言计算产品性能，最后将性能数据展示在网页上的整个日志采集及分析过程。本课题主要包含以下方面的内容：

1. Aria上日志的收集：日志收集主要通过Aria SDK实现，通过Tenant Token将日志与小娜联系起来，日志的内容通过Key-Value的方式插入；在此方面还包含日志的设计部分，为了使日志计算更加合理，设计了一系列措施保证数据完整性和准确性。
2. Cosmos上日志的计算：本系统采用Scope语言进行脚本的计算，通过计算日志之间的时间差和开始到结束日志数量差分别计算产品各个功能的延迟和成功率，延迟的计算方法不是采用简单的平均值，而是采用求分位数的方法求数据最低的75%、90%、95%三个数作为延迟。
3. 自动化工具实现：为了能自动的提交脚本并将计算结果导入到数据库中，本系统实现了定时提交脚本并导入数据，并且对提交脚本计算数据和导入数据库过程中可能出现的所有异常进行了相应处理，保证工具的健壮性和灵活性。
4. 数据库和网站实现：本系统使用SQL Server数据库，网站实现主要采用ASP.NET；网站将数据库中数据已可视化的图表曲线图的方式展示出来，可以直观地看到各个功能模块的性能参数。

## 论文的组织安排