# 一、项目名称：Citi PB核心交易系统log大数据分析与监控

## 二、立项依据：（技术合同的内容，技术创新点）

## **立项源由：**

由于花旗业务的特点，我们有好几个核心部署在不同的节点，每个节点的log相互独立，造成信息孤岛。我们想从各个核心系统提取交易信息来做分析去挖掘新的业务。由于纽约，伦敦，东京证券交易市场发送的交易数据并不固定。核心系统在处理大批量有高峰瓶颈，我们想了解核心系统处理交易数据的可靠性，和消耗时间，判断是否有崩溃的风险。所以我们急需一套系统能够把我们所有的核心系统监控起来，做风险预警。

## **技术内容：**

## Citi PB核心交易系统log大数据分析与监控（以下简称“系统”）是花旗集团（以下简称“本集团”） 基于大数据平台之上开始的一套log数据分析与监控风险预警系统。基础设施采用Hadoop平台，数据采集使用flume，消息中间件使用kafka，计算模型采用spark流式计算，外部存储采用mongo，报表展现使用zeppelin。

## 技术创新点

* + 1. 将spark架构在hadoop平台上保证系统的可靠性和稳定性。
    2. 弃用mapreduce采用spark计算，提高流式计算的效率。
    3. 使用kafka作为消息中间件，保证数据不丢失与吞吐量。
    4. 在各个核心系统的节点上采用flume实时监控，可以灵活的写插件提取客户需要的数据，提高可扩展性。
    5. 采用zepplin，用户可能灵活使用sql,mongo,spark,java等各种语言进行分析或生成报表。
    6. 在kafka中建立多个topic，可能灵活在系统上运行好几条业务线，便于扩展开发和维护。

**技术开发内容及形式：**

Flume集群从各个核心系统节点采取相关的log信息—>将采集的log信息发送到kafka集群—>spark集群从kafka中取出数据进行清洗，然后转化成标准的数据格式进行计算—>将加工好的计算结果存入mongo集群—>用户通过zepplin从mongo中取出数据进行分析与报表处理—>加工成报表展现给用户。

开发要求及目标：应达到的技术指标和参数

* 1. 实时并行的监控各个节点的log
  2. 高吞吐量（kafka保证）
  3. 高可靠性，failover机制（hadoop保证）
  4. 数据处理规模不少于600W
  5. 开发过程中使用git进行版本控制，JIRA进行开发进度追踪与管理