项目概述

本次Spark项目的展示包括多个电子商务相关的数据集分析，并通过SparkSQL进行处理，最终将分析结果通过积木报表进行可视化。团队成员分别从电子商务综合性及欺诈分析、京东平台冰箱销售数据分析以及化妆品电商数据分析三个方面进行展示。

需求分析Ⅰ

电子商务综合性及欺诈分析：

需求数据集：关于电子商务交易的全面信息，涵盖交易金额、支付方式、产品类别、客户信息等。

需求功能：分析交易行为和模式，发现潜在的欺诈行为。

目标：通过Spark进行实时数据处理，并利用积木报表进行可视化展示。

需求分析Ⅱ

京东平台冰箱销售数据分析：

需求数据集：京东2020年5月25日的冰箱订单数据，经过10%随机抽样和数据脱敏处理。

需求功能：统计冰箱销量、品牌销售总量、平均销售价格及各省份销量，并进行可视化展示。

需求分析Ⅲ

化妆品电商数据分析：

需求数据集：某化妆品电商十一月所有已注册用户的行为数据，包括浏览、加入购物车、移除购物车、购买等。

需求功能：分析销量前五的品牌、用户价值人数及占比、用户转换率、双十一期间的销售量变化。

目标：为平台和商家提供参考，通过RFM模型分析用户价值，制定更有效的市场策略。

技术可行性分析

Spark

Spark是一个强大的开源分布式计算系统，适合处理大规模数据集。

SparkSQL提供了对结构化数据的查询和分析能力，能够高效地进行数据处理和分析。

积木报表：

积木报表是一种灵活的报表工具，支持多种数据源和复杂的报表设计，适合用于数据的可视化展示。

Scala：

Scala是一种支持面向对象和函数式编程的语言，与Spark紧密集成，能够编写高效的Spark应用程序。

资源可行性分析

硬件资源：

需要配置集群环境以支持Spark的分布式计算。集群的规模和配置需要根据数据量和计算任务的复杂度进行调整。

软件资源：

需要安装和配置Spark、Scala、积木报表等软件工具。

数据源的获取和预处理需要相应的工具和脚本支持。

人力资源：

团队成员需要具备Spark和Scala的开发技能，并对数据分析和可视化有一定的经验。

风险分析

数据质量：

数据源的质量直接影响分析结果的准确性和可靠性。需要对数据进行清洗和预处理，确保数据的准确性。

性能瓶颈：

大规模数据处理可能会遇到性能瓶颈，需要优化Spark任务和集群配置，提高计算效率。

技术挑战：

Spark和Scala的开发需要一定的技术门槛，团队成员需要经过培训和学习，掌握相关技术。

总结

本次Spark项目通过对电子商务综合性及欺诈分析、京东平台冰箱销售数据分析以及化妆品电商数据分析，展示了Spark在大数据处理和分析方面的强大功能。通过SparkSQL进行数据处理，并利用积木报表进行可视化，能够满足项目的需求。项目在硬件、软件和人力资源方面具备一定的可行性，但需要注意数据质量和性能优化的风险。整体来看，项目具有较高的可行性和实施价值。