基于 TPC-H 数据集的可视化数据分析

作业总体要求：

基于 TPC-H 数据集，模拟淘宝数据分析场景，综合使用数据库课程知识撰写一篇数据分析报告，通过数据库管理、数据分析及数据可视化功能综合展示数据内部的规律。

报告技术脉络：

- SQL Server 数据库：数据存储、管理、数据预处理、数据处理、视图管理
- Tableau&PowerBI：数据可视化分析工具，通过多维、多层次方法全面、综合展现数据规律与趋势
- SQL Server 数据库分析：构建复杂查询任务，处理复杂数据分析问题，构建复杂数据分析视图
- Python：数据挖掘及深入分析，探索数据内部规律及预测，综合运用 SQL 语言构建分析数据集、存储分析结果、查看分析结果，以及使用数据可视化工具展现数据挖掘结论

1. 版式：

要求参考毕业论文或者项目报告的格式完成一篇 TPC-H 数据分析技术文档，要求设置标题层次，通过【视图】-【导航窗格】可以显示文档结构图，通过目录定位。

文档中通过屏幕抓图说明实现的结果，要求设置图题。

要求文档格式美观，层次清晰，易读易懂。

2. 数据集

使用 SQL Server 中创建 TPC-H 数据库导入或 DirectQuery。在 PowerBI 或 Tableau 中创建表关系，修改维-度量属性，确定用于分析的维属性，用于聚集计算的度量属性及设置维中的层次。说明维、层次、度量的设置方法。

3. 主题分析

分析 TPC-H 数据集的结构和数据特征，从 customer、supplier、part 三个维度来分析销售数据在不同视角，不同分析粒度上的数据规律与特征，综合运用数据可视化组件展现数据特点并加以说明。

4.仪表板设计

设计一个全局的仪表板（报表）页，设计面向 TPC-H 数据集的分析方案，通过多个图表控件设计一个能够良好展现数据宏观、微观及特定分析主题的综合数据视图，说明仪表板的使用及所展现的数据分析结果。

5.SQL 查询

通过 SQL 语言构建常规多维分析无法解决的复杂分析任务，设计 3 个基于派生查询的复杂查询任务，实现面向现实应用需求、基于原始数据再加工的分析任务，并给出查询优化方案。

6.Python 分析

基于 Python(独立或 SQL Server 内置)实现对 TPC-H 数据的深度分析 deep-analysis，构建一个分析主题(如客户价值分析、购物篮分析等综合分析任务)，围绕分析主题综合运用 SQL 语言进行数据组织、数据预处理理、构建分析模型、运行分析任务、分析数据挖掘结果并通过数据可视化工具展现数据挖掘结论。

报告的目标是模拟企业数据分析场景中的综合数据分析任务，学习如何以数据为中心构建数据分析框架，如何按照数据分析需求组织数据、数据预处理，如何通过数据可视化工具快速展现数据的基本规律和多维度分析结果及结论，如何通过 SQL 语言处理复杂查询任务，为数据可视化及数据挖掘构建分析数据视图，如何通过 Python 与数据库及数据可视化技术

的结合深入挖掘数据内在的规律与价值。

课程报告的目的一方面考查大家对数据分析处理技术框架所涉及的一系列数据管理与数据处理工具能否熟练运用，另一方面也通过 TPC-H 数据分析案例增强大家对企业数据分析能力的训练，让报告变为企业工作的实战模拟，为未来的职业做一个好的数据分析工作模板。

大家可以将 TPC-H 数据模拟为双 11 的销售数据，模拟双 11 过后的数据分析任务。

祝好

张延松