基于 Presto 的华盛顿州电动汽车数据分析

小组成员	
	黄勖-22920212204392
	曾鸿勇-22920212204488
	石宇昊-22920212204434
	邹慧-22920212204513

目录

基于 Presto 的华盛顿州电动汽车数据分析 1

一、项目背景 2

二、需求分析 2

三、逻辑设计 4

四、物理设计 4

五、系统实现 7

1.系统搭建 7

2.数据源 12

3.ETL 过程 15

3.1 数据抽取 15

3.2 数据转换 17

3.3 数据装载 18

六、数据可视化及数据分析 21

七、总结 27

一、项目背景

我们搜集了 2010-2023 电动汽车车辆销售与所有权变更活动数据集，这个数据集记录了华盛顿州电动汽车约 85 万所有权变更和车辆注册交易。该数据集详细记录产权的变更情况，以及授权车辆在华盛顿州公共市场上进行的交易。这些记录提供了有关电动汽车所有权和使用模式的关键信息，有助于了解该州电动汽车市场的趋势和普及率。这些数据对于政策制定者和行业利益相关者具有重要价值，希望我们的分析能够为电动汽车市场的发展动态提供帮助，评估市场变化的影响，并制定支持电动出行在华盛顿州可持续发展的战略。 可以帮助新能源汽车投资人考虑如何投资使利益最大化。

二、需求分析

为了更好地理解和利用我们收集的电动汽车统计数据，我们明确以下需求：

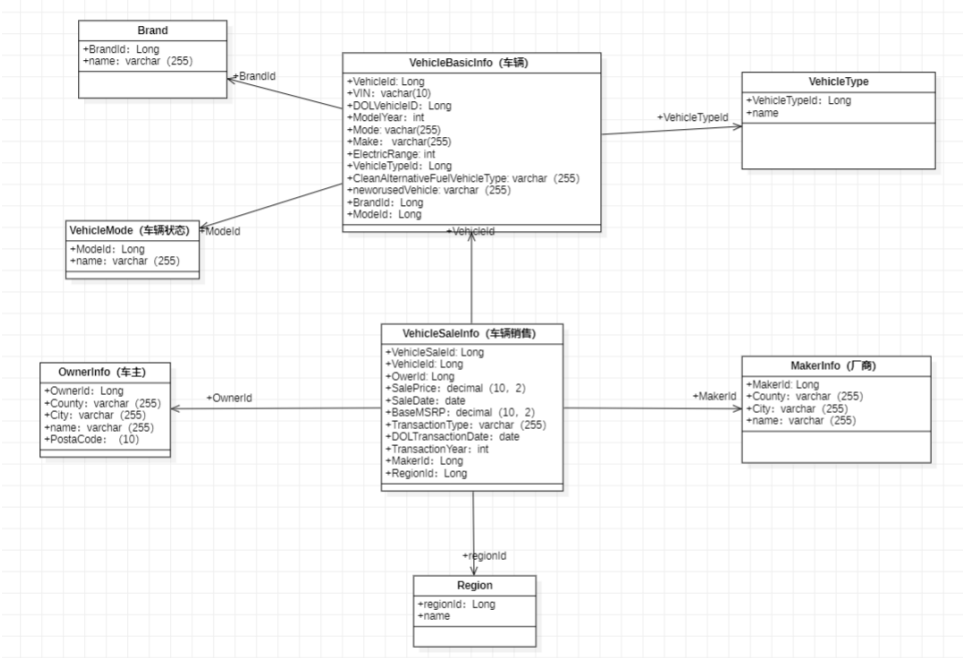
1. **数据提取需求：**
    - 实现定期自动提取电动汽车车辆登记与所有权变更活动数据集的更新。
    - 支持增量加载，以最小化数据提取的时间和资源成本。
    - 考虑对原始数据的清洗和转换，以适应数据仓库的结构和标准。
2. **数据模型设计需求：**
    - 创建清晰的数据模型，包括维度和事实表，以支持复杂的查询和报告需求。
    - 包括车辆信息、所有权变更事实、注册活动事实等维度和事实表。
    - 考虑与其他相关数据集的集成，以提供更全面的分析。
3. **数据质量需求：**
    - 实施数据质量控制策略，包括验证关键字段的准确性和完整性。
    - 设定数据质量指标，监测和报告数据质量问题。
    - 提供数据纠错和清洗的机制，确保数据仓库中的数据是可信的。
4. **性能和可扩展性需求：**
    - 优化查询性能，确保用户能够在合理的时间内获取结果。
    - 考虑分区表、索引等技术手段，以提高数据仓库的查询效率。
    - 预测未来数据增长，确保数据仓库具有足够的可扩展性。

5. **安全和权限需求：**
    - 实施访问控制，确保只有授权用户可以访问敏感信息。
    - 对数据进行脱敏或加密，以保护用户隐私。
    - 遵循合规性要求，特别是关于个人身份信息和敏感数据的法规。
6. **报表和分析需求：**
    - 提供用户友好的报表和可视化工具，以便用户能够轻松理解和分析数据。
    - 支持多维分析，允许用户针对不同维度进行深入挖掘。
    - 提供自定义报表和查询功能，以满足各种业务需求。
7. **用户培训和支持：**
    - 提供项目设计详细文档，确保阅读人员能够充分利用数据仓库的功能。
    - 设立支持渠道，为用户提供技术支持、解决问题。
8. **可追溯性和审计需求：**
    - 记录数据变更和访问日志，以实现数据的可追溯性。
    - 实施审计机制，以满足合规性和法规要求。

以上需求分析旨在确保数据仓库项目能够有效地满足用户和业务的需求，并提供高质量、安全和可扩展的数据分析平台。

三、逻辑设计

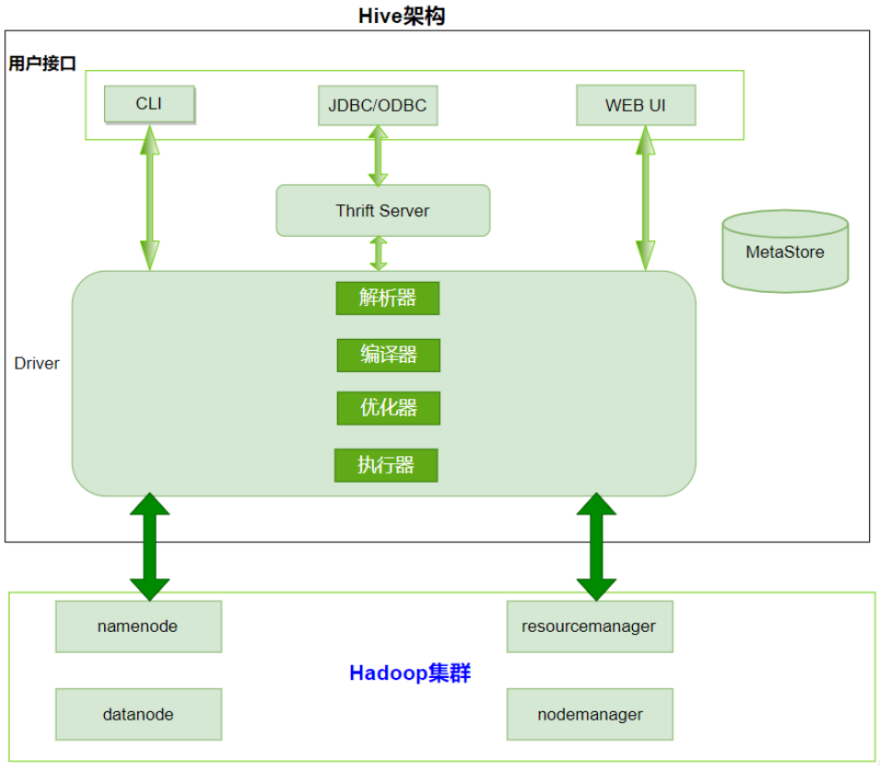
事实表和维度表



这一部分我们采用了**雪花模型**，以“车辆销售”作为事实表，基于实际的需求分析构建了如图所示的七个维度表，其中由于车辆信息中我们又分析了多个维度，因此在此部分又分成了多个维度表。

四、物理设计

我们的数据仓库物理结构的解决方案基于 Hadoop、Hive 和 Presto，形成了一个强大的大数据解决方案。Hadoop 作为底层分布式存储和计算框架，提供高可靠性和可扩展性，存储大规模数据。Hive 作为数据仓库的元数据存储和查询引擎，通过 HQL（Hive Query Language）将结构化查询转化为 MapReduce 任务，实现高层次的 SQL-like 查询。Presto 则作为交互式查询引擎，支持快速查询和分析，通过连接 Hive 等数据源直接执行 SQL 查询，提供了更低延迟的数据分析体验。这一集成架构为我们小组的数据仓库提供了高效、灵活和可扩展的物理结构，支持复杂的大数据分析和查询需求。



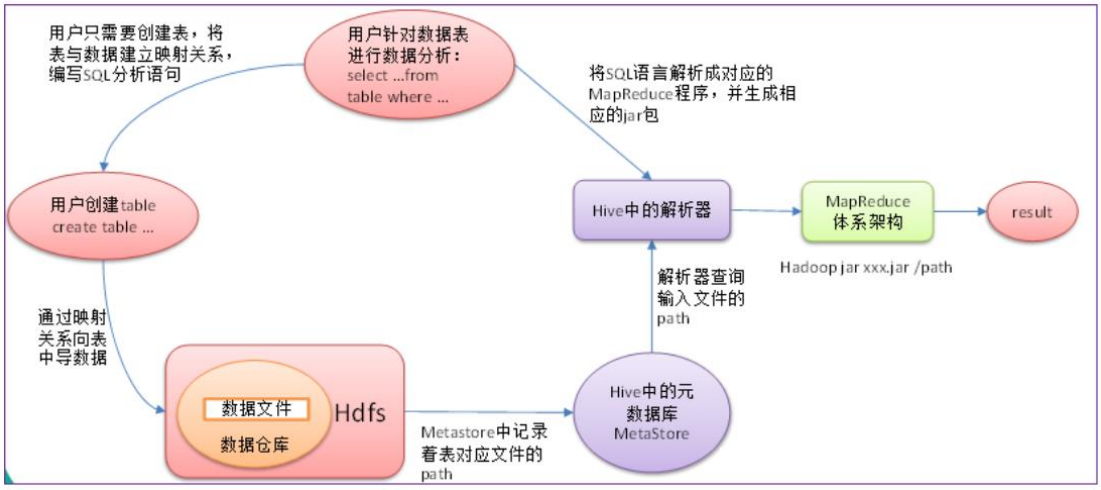
每层的组件详细介绍如下：

Hadoop：

Hadoop 是一个开源的分布式存储和计算框架，它提供了处理大规模数据的能力。Hadoop 的内核组件包括 Hadoop Distributed File System（HDFS）和 MapReduce。在物理设计层面，Hadoop 采用分布式存储模型，将数据划分为多个块并存储在不同的节点上，以实现高度的可扩展性和容错性。物理设计方面的考虑包括数据块的复制、分片、数据压缩等。

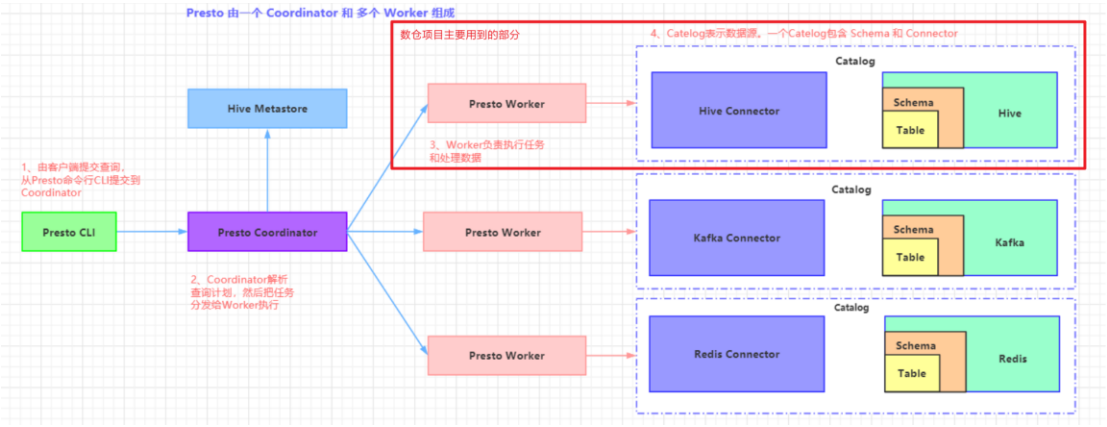
Hive：

Hive 是一个基于 Hadoop 的数据仓库工具，它提供了类似 SQL 的查询语言（HiveQL），允许用户通过 SQL 方式查询和分析存储在 Hadoop 中的数据。在物理设计上，Hive 将查询转化为一系列的 MapReduce 作业，这些作业在 Hadoop 集群上并行执行。Hive 使用元数据来描述数据的结构和位置，通过表的分区、桶等方式进行数据组织和优化查询性能。



Presto：

Presto 是一个高性能的分布式 SQL 查询引擎，用于查询大规模分布式数据。与 Hive 不同，Presto 不使用 MapReduce，而是通过使用自己的执行引擎来直接查询数据。在物理设计上，Presto 采用内存计算，数据以列存储的形式进行存储和处理，以提高查询性能。Presto 支持多种数据源，包括 Hive、关系型数据库、NoSQL 数据库等。



为了确保设计的表在物理层面上能够有效地支持数据存储和查询操作，以下是一些我们小组初步设计的物理设计要求：

1. **索引设计：** 对于经常用于查询条件的列，考虑创建索引以提高查询性能。例如，在 VehicleBasicInfo 表中，VIN 是主键，已经具有索引。在其他表中，还有对经常用于筛选或连接的列创建索引，如 SaleDate、TransactionType 等。

2. **数据类型选择：** 选择适当的数据类型以最小化存储空间并提高性能。例如，在 VehicleBasicInfo 表中，对于表示金额的列，选择了 DECIMAL(10, 2)。确保其他列的数据类型也符合实际需求，避免使用过大或不必要的数据类型。

3. **外键关系：** 外键关系已经被正确地定义，这有助于确保数据的一致性。确保外键列上有适当的索引，以提高连接操作的性能。

4. **分区和分桶：** 考虑在表的设计中使用分区和分桶，以便更有效地管理和查询大量数据。这在处理大型数据集时可以提高查询性能。

5. **统计信息和优化：** 定期更新数据库统计信息，以确保查询优化器能够制定有效的执行计划。这对于复杂查询和大型数据集尤为重要。

6. **表空间和存储设置：** 确保数据库表和索引存储在适当的表空间中，并考虑合理的存储设置，如数据文档大小和增长参数。

7. **备份和恢复策略：** 实施合适的数据库备份和恢复策略，以保障数据的可靠性和安全性。

8. **数据质量和一致性：** 确保数据的准确性和一致性，通过实施约束、触发器等机制来强制执行业务规则。

这些建议可以根据具体情况进行调整，确保表的设计在物理层面上既能够满足性能要求，又能够保持数据的一致性和可靠性。

五、系统实现

1.系统搭建

我们的数据仓库系统的搭建是一个综合性的过程，它牵涉到硬件和软件层面的配置。以下是系统搭建的主要组成部分：

硬件配置：

- **操作系统：** 选择了 CentOS 作为操作系统，它是一种广泛用于企业环境的 Linux 发行版，CentOS 作为操作系统和虚拟机的规格都是合理的选择，提供了稳定性和足够的资源来支持数据仓库的运行。
- **硬件规格：** 每台虚拟机配置了 4GB 的内存，共有 3 台虚拟机。此外，每台虚拟机还有 50GB 的硬盘空间，用于存储系统和应用程序的数据。

软件配置：

- **数仓平台：** 采用 Hadoop 和 Hive 作为数仓平台，Hadoop 用于分布式存储

和处理大规模数据，而 Hive 提供了 SQL 接口，使用户能够通过 SQL 查询分布式存储的数据。

- **查询引擎：** Presto 充当了交互式查询引擎的角色。Presto 能够连接 Hive 等数据源，通过执行 SQL 查询，提供了更低延迟的数据分析体验。用户可以通过 Presto 轻松地进行复杂查询、探索性分析和交互式数据探查。
- **ETL 工具：** 选择了 Kettle 作为 ETL（抽取、转换、加载）工具，它能够方便地处理数据的提取、清洗和加载，是数据仓库中常用的工具之一。
- **数据可视化工具：** 使用 Superset 作为数据可视化工具，Superset 能够通过简单的配置和交互性地创建丰富的数据可视化报表，支持多种数据源。

硬件+软件搭配架构：

- 系统采用了分布式架构，Hadoop 提供了分布式存储和计算能力。
- ETL 工具 Kettle 负责将数据从源系统提取到数仓中，并在此过程中进行必要的转换和清洗。
- 数据可视化工具 Superset 通过连接到 Hive 或其他数据源，实现对数据仓库中数据的可视化展示。

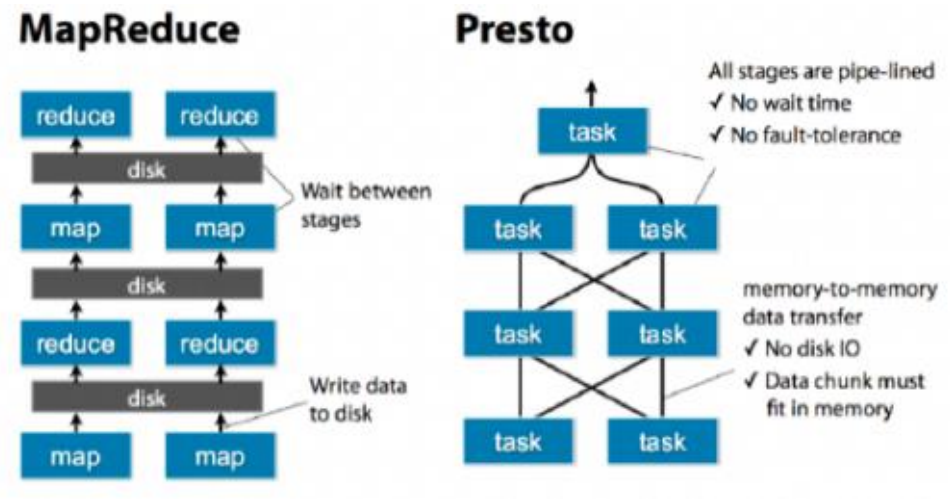
整个系统架构体系充分利用了分布式计算和存储的优势，使得大规模数据的处理和分析变得高效而可行。同时，ETL 工具和数据可视化工具的引入使得数据流程更加流畅，使用户能够更轻松地管理和分析数据。系统搭建的成功将为企业提供强大的数据处理和分析能力，支持业务决策和洞察。

Presto 介绍与配置：

（1）Presto 介绍

Presto 是一款 Facebook 开源的 MPP 架构的 OLAP 查询引擎，可针对不同数据源执行大容量数据集的一款分布式 SQL 执行引擎，数据量支持 GB 到 PB 字节，主要用来处理秒级查询的场景。Presto 本身并不存储数据，但是可以接入多种数据源，并且支持跨数据源的级联查询，而且基于内存运算，速度很快，实时性高。虽然 Presto 可以解析 SQL，但它不是一个标准的数据库。不是 MySQL、Oracle 的代替品，也不能用来处理在线事务(OLTP)。

适合 PB 级海量数据复杂分析，交互式 SQL 查询，支持跨数据源进行数据查询和分析。不像 hive，只能从 hdfs 中读取数据。



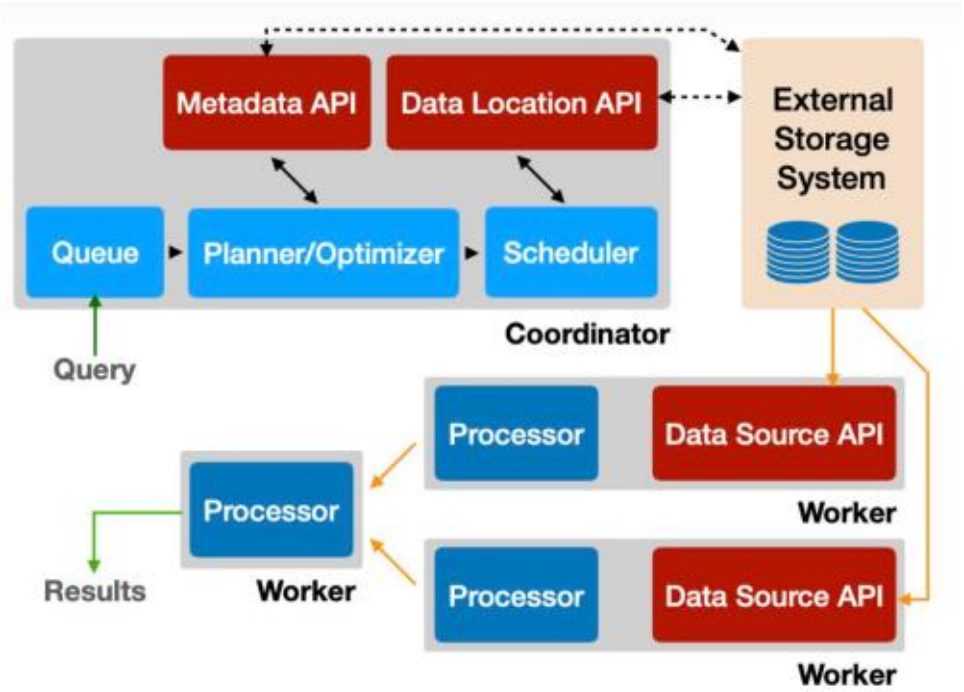
（2）Presto 与 Hive 对比

hive 是一个数据仓库（有 hive 表），是一个交互式比较弱的查询引擎，交互能力没有 presto 那么强，而且只能访问 hdfs 的数据，数据源比较单一，Presto 是一个交互式查询引擎，可以在很短的时间内返回查询结果，秒级，分钟级，能访问很多数据源 hive 在查询 100Gb 级别的数据时，消耗时间已经是分钟级了。

presto 是取代不了 hive 的，因为 presto 全部的数据都是在内存中，限制了在内存中的数据集大小，比如多个大表的 join，这些大表是不能完全放进内存的，所以 presto 不适合用在多个大表的 join。

实际应用中，对于在 presto 的查询是有一定规定条件，查询过大，会占用整个集群的资源，这会导致你后续的查询是没有资源的，我们理想的交互应该是实时的，速度越快越好。

Presto 通过使用分布式查询，可以快速高效的完成海量数据的查询。如果你需要处理 TB 或者 PB 级别的数据，那么你可能更希望借助于 Hadoop 的 HDFS 来完成这些数据的处理。作为 Hive 和 Pig（Hive 和 Pig 都是通过 MapReduce 的管道流来完成 HDFS 数据的查询）的替代者，Presto 不仅可以访问 HDFS，也可以操作不同的数据源，比如 mysql。



（3）Presto 环境配置

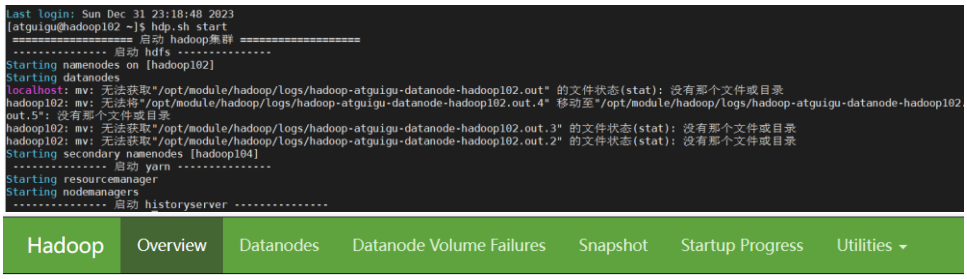
Presto 通过使用分布式查询，可以快速高效的完成海量数据的查询。如果你需要处理 TB 或者 PB 级在安装目录中创建一个 etc 目录，此目录下将会包含以下配置文件：

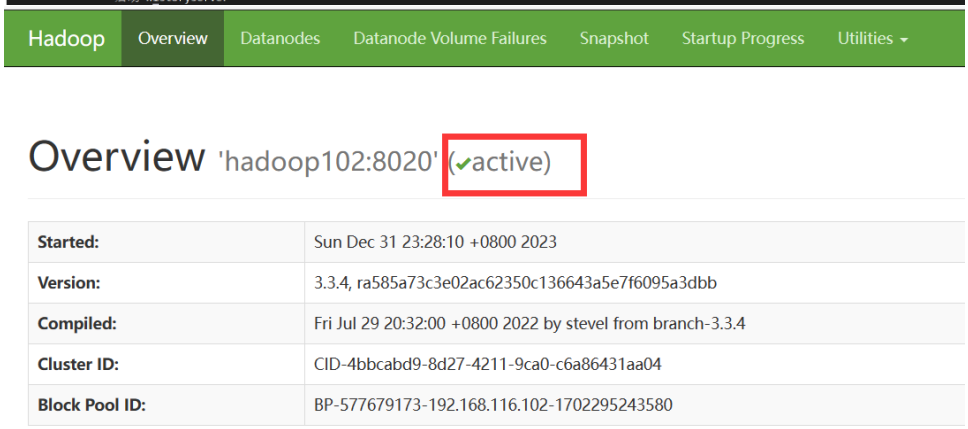
- node.properties: 每个节点的环境配置
- jvm.config: JVM 的命令行选项
- config.properties: Presto Server 的配置项
- catalog/hive.properties: 数据源连接器的配置
- Etc/node.properties: 每个节点的特定配置，一个节点指的是服务器上 Presto 的单个已安装实例。
- JVM 虚拟机配置: etc/jvm.config，包含用于启动 Java 虚拟机的命令行选项列表。文件的格式是选项列表，每行一个。不能使用空格或其他特殊字符。
- Presto 服务配置: etc/config.properties，包含 Presto 服务器的配置。Presto 服务分为三种角色：coordinator、worker、coordinator&worker。每个 Presto 服务都可以充当 coordinator 和 worker，但是独立出一台服务器专用于 coordinator 协调工作将在较大的群集上提供最佳性能。

环境启动与测试：

- **Hadoop 集群启动：**Hadoop 采用分布式存储方式，在 Master 主节点机启动 Hadoop 集群，其他配置环境（presto 除外）也仅需在此节点启动

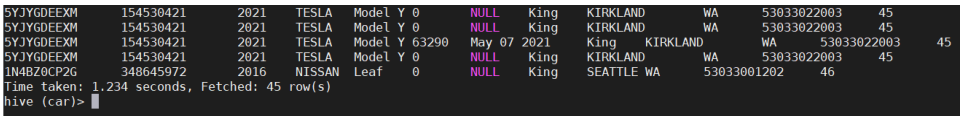
hdp.sh start(master)命令启动Hadoop，可以在控制台[hadoop控制台](#)，查看Hadoop启动情况





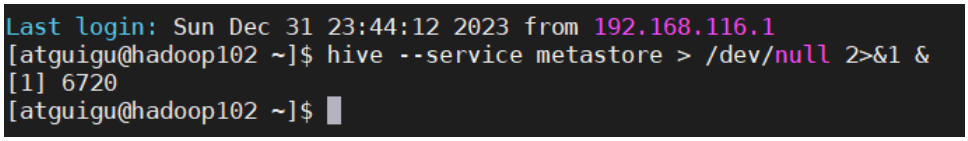
- **Hive 数据仓库启动与测试**：启动 Hive 数据仓库，尝试读取数据仓库，查看数据仓库是否成功配置

测试 select 语句，并记录相关查询时间，方便后期与配置 presto 的查询速度进行对比

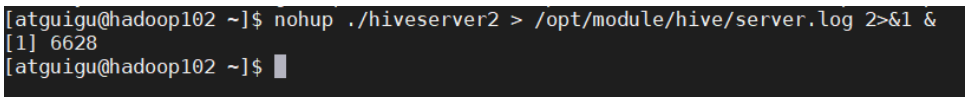


- **Hiveserver Hiveserver2 启动**：采用后台启动的方式，启动 hiveserver 方便后续配置使用

启动 hive 元数据存储，presto 需要使用此 hive 元数据进行操作



启动 hiveserver，以便 presto、superset 使用



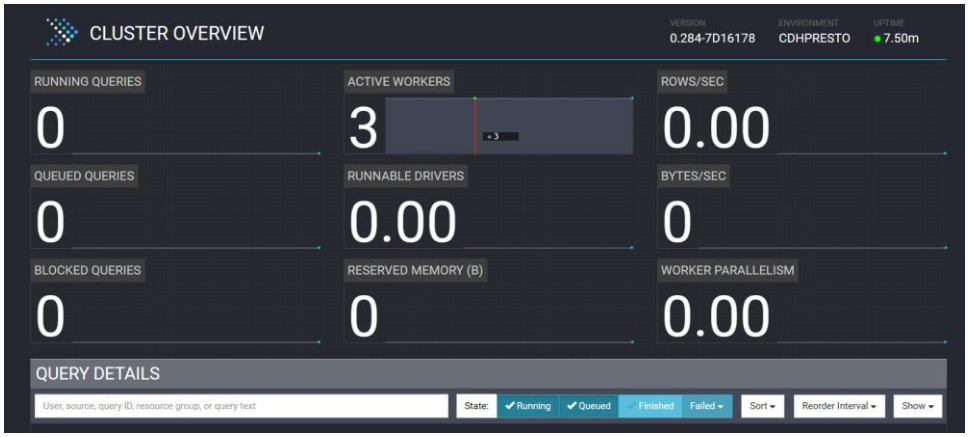
- **Presto 启动与对比测试：**presto 启动与上文不同，它需要在三个节点均启动

才可供后续查询使用

./start_presto.sh(全节点)命令启动 presto 服务

```
Last login: Sun Dec 31 23:44:38 2023 from 192.168.116.1
[atguigu@hadoop102 ~]$ ./start_presto.sh
Started as 6938
[atguigu@hadoop102 ~]$
```

可以在控制台 presto 控制台，查看 Hadoop 启动情况



启动 presto 进行测试，采取与 hive 相同的数据库进行 select 测试，可以 presto 的读取速度，但第一次无法体现 prestoc 采用内存机制来提高读取速度的优越性

```
Last login: Sun Dec 31 23:44:38 2023 from 192.168.116.1
[atguigu@hadoop102 ~]$ ./start_presto.sh
Started as 6938
[atguigu@hadoop102 ~]$ presto --server 192.168.116.102:8090 --catalog hive --schema car
presto:car> select* from cars;
```

可以看到 presto 第二次查询已经达到了 65rows/s，已经超过了 hive 的读取速度

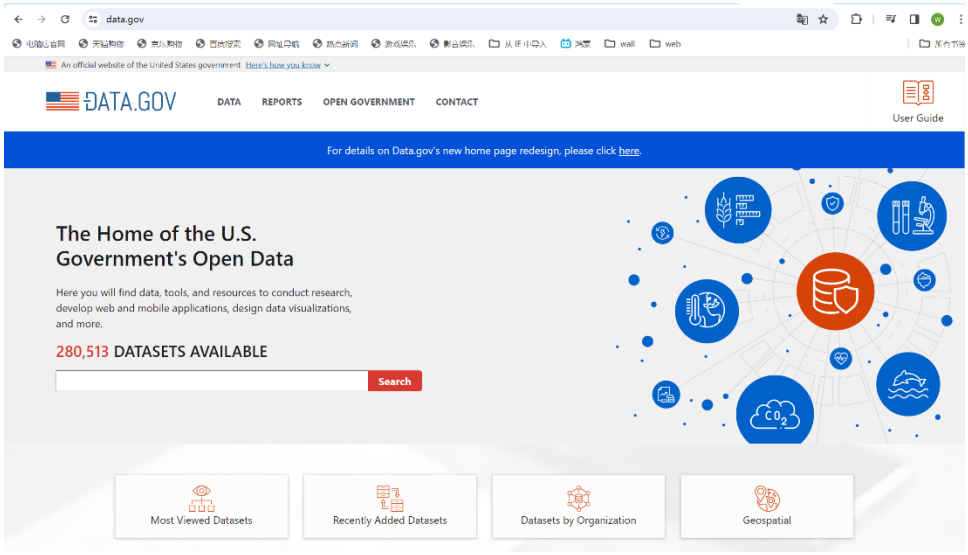
```
Query 20231231_155615_00003_st67n, FINISHED, 3 nodes
Splits: 61 total, 61 done (100.00%)
[Latency: client-side: 0:01, server-side: 0:01] [45 rows, 3.64KB] [65 rows/s, 5.3KB/s]
```

而 presto 第三次查询速度已经达到了 93rows/s,高速的读取速度，提高了因为 hive 限制而引起的 superset 数据分析以及可视化的速度

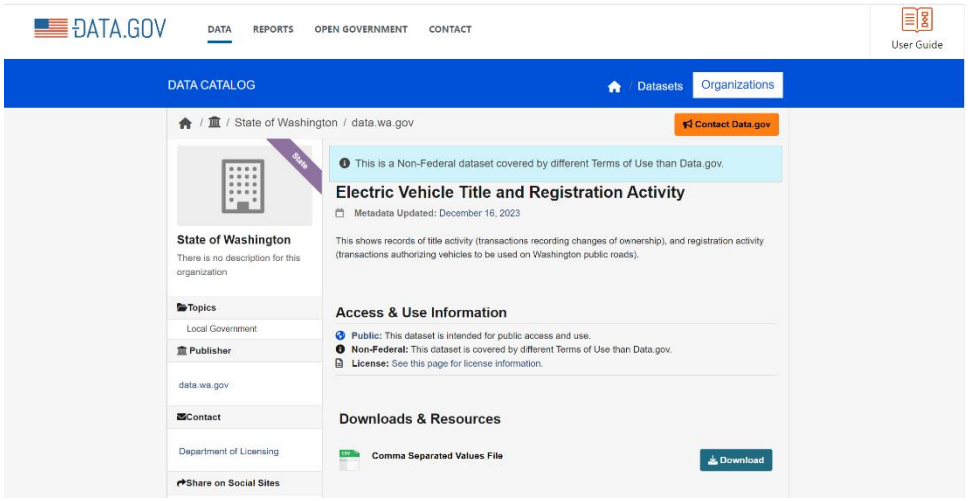
```
Query 20231231_155624_00004_st67n, FINISHED, 3 nodes
Splits: 61 total, 61 done (100.00%)
[Latency: client-side: 495ms, server-side: 479ms] [45 rows, 3.65KB] [93 rows/s, 7.62KB/s]
```

2.数据源

在本次实验中，数据来源于 美国政府官方开放数据网站 ；



我们选择 **Electric Vehicle Title and Registration Activity** 作为本次大作业的元数据；



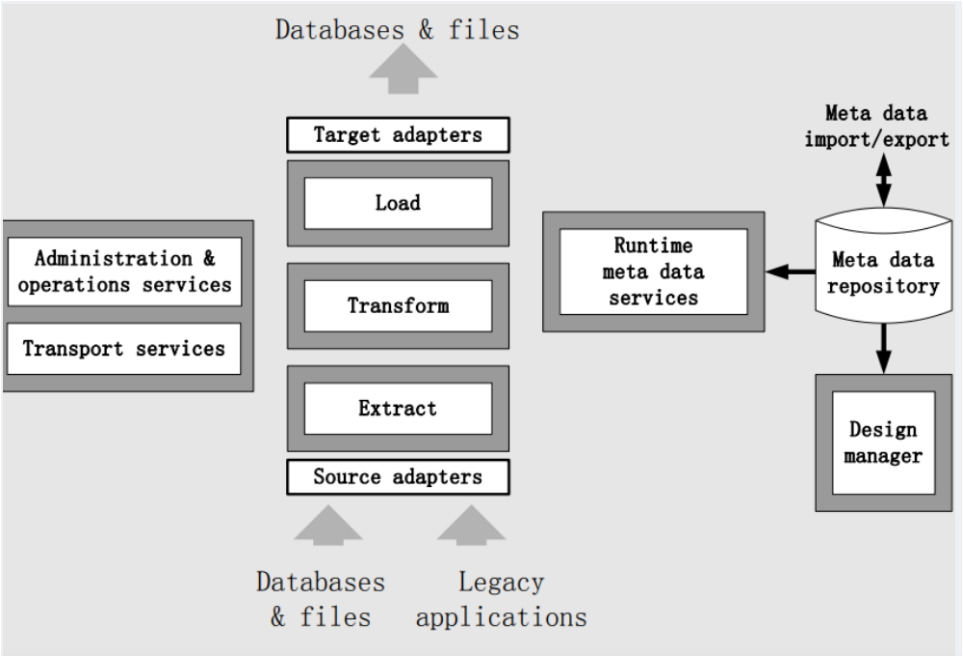
该 csv 文件中共有 35 个字段，以下是所有字段的内容：

字段英文名	解释
Clean Alternative Fuel Vehicle Type	指示车辆的清洁替代燃料类型，可能包括电动、混合动力等
VIN (1-10)	车辆识别号，由 10 个字符组成
DOL Vehicle ID	部门车辆标识符，用于唯一标识每辆车
Model Year	车辆制造年份
Make	车辆制造商
Model	车型
Vehicle Primary Use	车辆的主要用途，例如个人用车、商务用车等
Electric Range	电动车辆的续航里程
Odometer Reading	车辆当前的里程数

Odometer Code	里程计代码，描述里程数的单位（英里或公里）
New or Used Vehicle	指示车辆是新车还是二手车
Sale Price	车辆的销售价格
Sale Date	车辆销售日期
Base MSRP	车辆的制造商建议零售价格
Transaction Type	交易类型，例如购买、租赁等
DOL Transaction Date	部门交易日期
Transaction Year	交易年份
County	车主所在的县
City	车主所在的城市
State of Residence	车主的居住州
Postal Code	车主的邮政编码
2015 HB 2778 Exemption Eligibility	2015 年的一个法案（HB 2778）的豁免资格
2019 HB 2042 Clean Alternative Fuel Vehicle (CAFV) Eligibility	2019 年的一个法案（HB 2042）关于清洁替代燃料车辆资格
Meets 2019 HB 2042 Electric Range Requirement	是否符合 2019 HB 2042 的电动车续航要求
Meets 2019 HB 2042 Sale Date Requirement	是否符合 2019 HB 2042 的销售日期要求
Meets 2019 HB 2042 Sale Price/Value Requirement	是否符合 2019 HB 2042 的销售价格/价值要求
2019 HB 2042: Battery Range Requirement	2019 HB 2042 的电池续航要求
2019 HB 2042: Purchase Date Requirement	2019 HB 2042 的购车日期要求
2019 HB 2042: Sale Price/Value Requirement	2019 HB 2042 的销售价格/价值要求
Electric Vehicle Fee Paid	支付的电动车辆费用
Transportation Electrification Fee Paid	支付的交通电气化费用
Hybrid Vehicle Electrification Fee Paid	支付的混合动力车辆电气化费用
2020 Census Tract	2020 年人口普查小区
Legislative District	所在的立法区
Electric Utility	使用的电力公用事业

*其中与日期有关的数据粒度为**日级别粒度**

3.ETL 过程



3.1 数据抽取

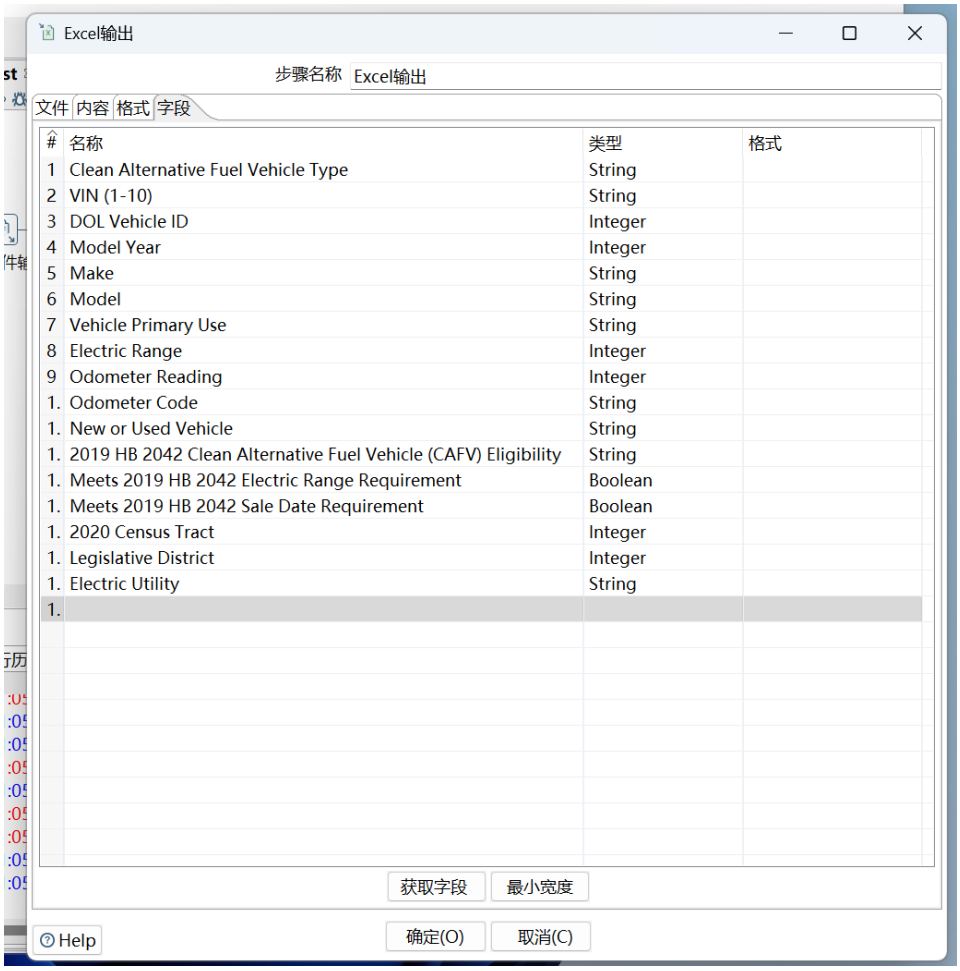
数据抽取指的是从不同的网络、不同的操作平台、不同的数据库和数据格式、不同的应用中抽取数据的过程。目标源可能包括 ERP、CRM 和其他企业系统，以及来自第三方源的数据。

从数据源中抽取满足多个维度数据，这里的维度数据包括：车辆销售地区维度表、厂商维度表、车辆品牌维度表，车辆状态维度表、车辆类型维度表等。

- 车辆类型维度表记录了车辆的新能源各类车型；
- 车辆销售地区维度表记录了车辆销售分布的各个地区；
- 厂商维度表记录了车辆销售中出现的厂商；
- 车辆状态维度表记录了车辆的状态情况；
- 品牌维度表记录了车辆的品牌信息。

使用 Kettle 作为 ETL 工具，新建一个转换：

获取我们需要的字段形成一个 excel 文件：



3.2 数据转换

数据转换实际上还包含了数据清洗的工作，需要根据业务规则对异常数据进行清洗，主要将不完整数据、错误数据、重复数据进行处理，保证后续分析结果的准确性。

数据转换就是处理抽取上来的数据中存在的不一致的过程。数据转换一般包括两类：第一类：数据名称及格式的统一，即数据粒度转换、商务规则计算以及统一的命名、数据格式、计量单位等；第二类：数据仓库中存在源数据库中可能不存在的数据，因此需要进行字段的组合、分割或计算。

我们实验中的 Electric_Vehicle_Title_and_Registration_Activity.csv 文件存在数据项内容为空、数据项格式不规范、存在重复列等问题，考虑使用 python 程序进行基本数据清洗活动。

①重复值处理：Electric_Vehicle_Title_and_Registration_Activity.csv 文件中存在部分重复行，影响数据分析的精确度，需要进行去重操作

```
#step2:数据去重
df = df.drop_duplicates()
```

②数据标准：“DOL Transaction Date” 、“Sale Date”此两列日期格式为“June 14 2023”，而在 Superset 可视化中对日期格式的要求比较严格，应为“2023-6-14”；特别的“Sale Date”列存在空值，将空值替换为默认日期“1999-9-9”

```
import pandas as pd

# 读取CSV文件
file_path = r"E:\Electric_Vehicle_Title_and_Registration_Activity.csv"
df = pd.read_csv(file_path)

#step1:日期格式转换

# 将 'DOL Transaction Date' 列转换为日期格式
df['DOL Transaction Date'] = pd.to_datetime(df['DOL Transaction Date'], format='%B %d %Y')

# 将 'Sale Date' 列转换为日期格式
df['Sale Date'] = pd.to_datetime(df['Sale Date'], format='%B %d %Y', errors='coerce')

# 将日期格式转换为 'YYYY-MM-DD'，同时处理空值
df['DOL Transaction Date'] = df['DOL Transaction Date'].dt.strftime('%Y-%m-%d')
df['Sale Date'] = df['Sale Date'].dt.strftime('%Y-%m-%d').fillna('1999-09-09')
```

③空值处理： Electric_Vehicle_Title_and_Registration_Activity.csv 文件中存在的空值问题，统一将其替换为“unkwon”字符串格式

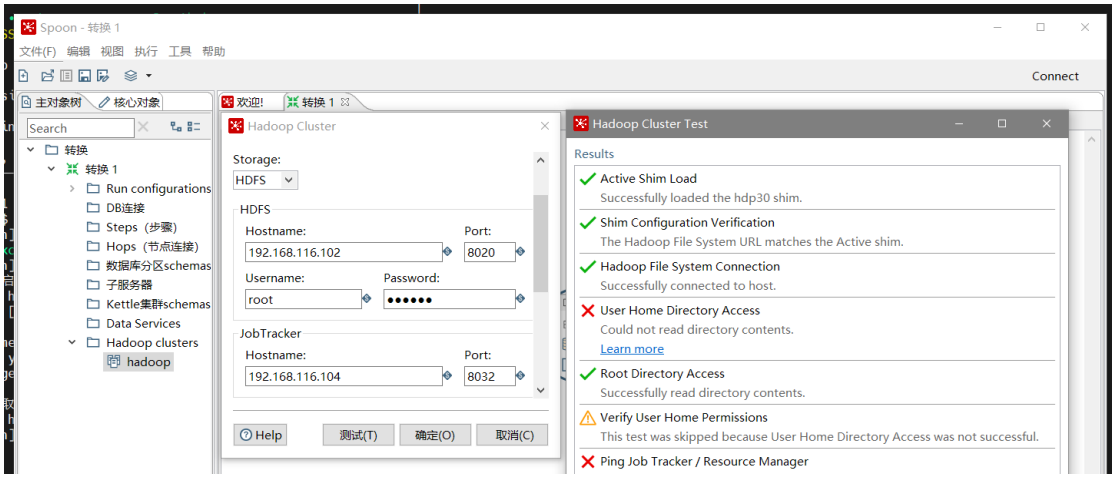
```
#step3:空值处理
# 将 '2019 HB 2042: Battery Range Requirement' 列中的空值替换为 'unknown'
df['2019 HB 2042: Battery Range Requirement'] = df['2019 HB 2042: Battery Range Requirement'].fillna('unknown')

# 将 'Transportation Electrification Fee Paid' 列中的空值替换为 'unknown'
df['Transportation Electrification Fee Paid'] = df['Transportation Electrification Fee Paid'].fillna('unknown')

# 将 'Hybrid Vehicle Electrification Fee Paid' 列中的空值替换为 'unknown'
df['Hybrid Vehicle Electrification Fee Paid'] = df['Hybrid Vehicle Electrification Fee Paid'].fillna('unknown')


output_file_path = r"E:\cleaned_data.csv"
df.to_csv(output_file_path, index=False)

print("已完成etl")
```

3.3 数据装载

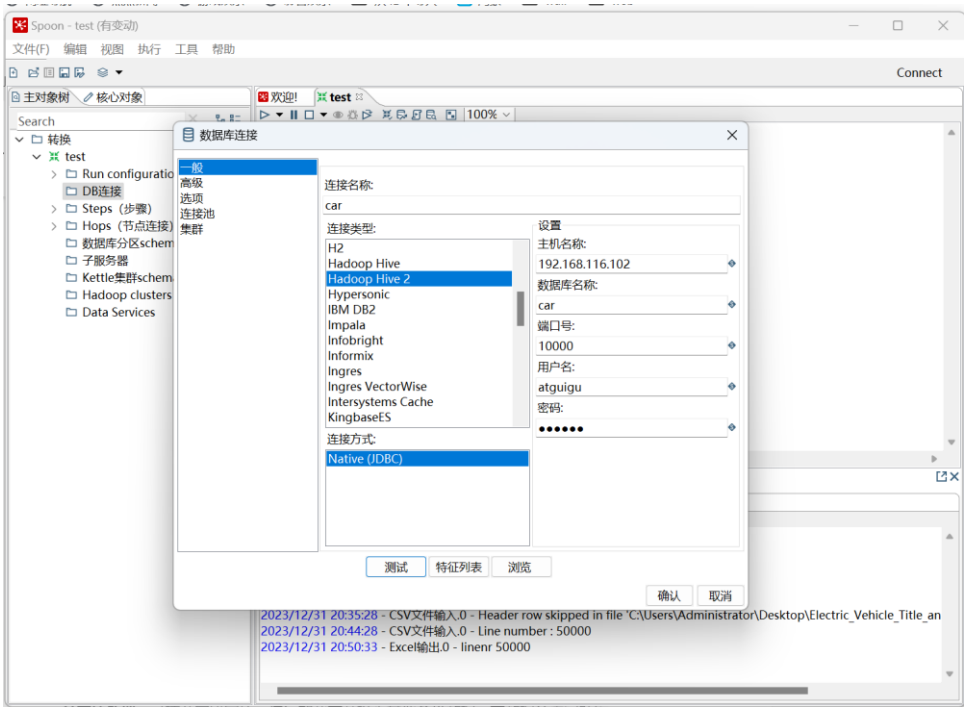
数据加载的主要任务是将经过清洗后的干净的数据集按照物理数据模型定义的表结构装入目标数据仓库的数据表中，如果是全量方式则采用 LOAD 方式，如果是增量则根据业务规则 MERGE 进数据库，并允许人工干预，以及提供强大的错误报告、系统日志、数据备份与恢复功能。整个操作过程往往要跨网络、跨操作平台。

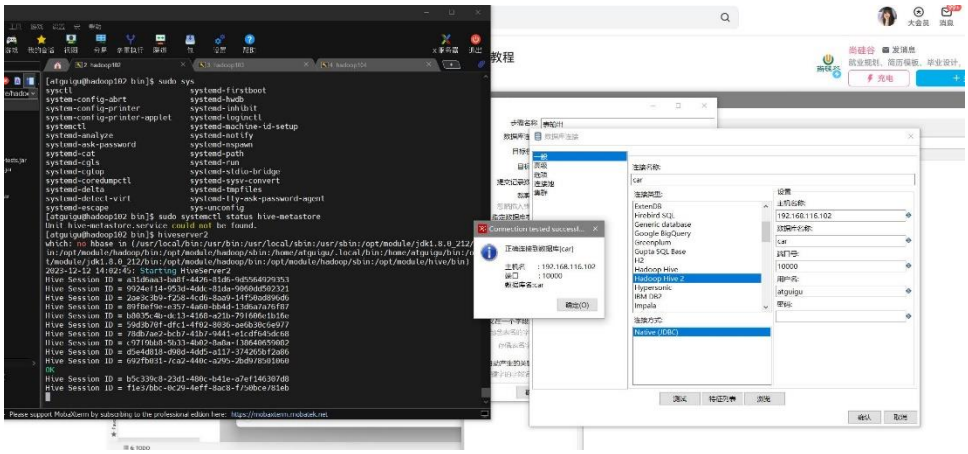
在我们的实际工作中，使用了 hive + presto，加速数据的读写。

首先连接 hadoop cluster：

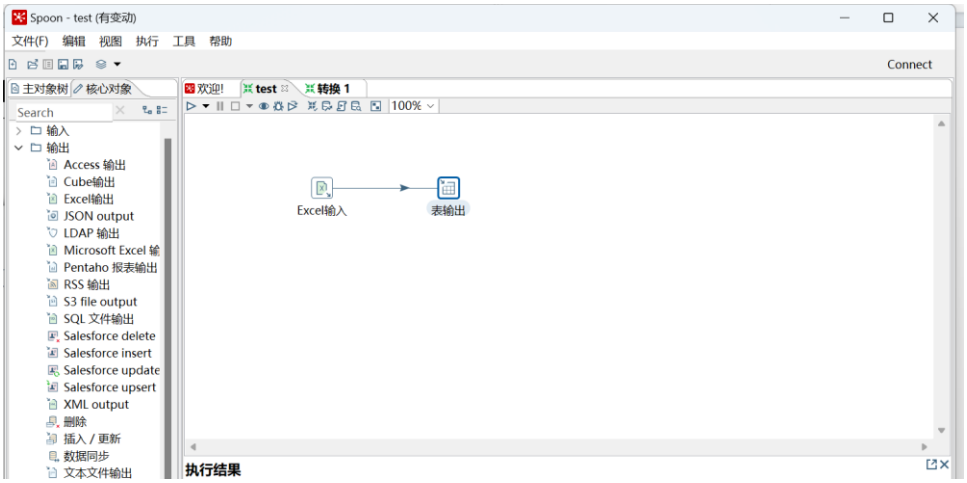


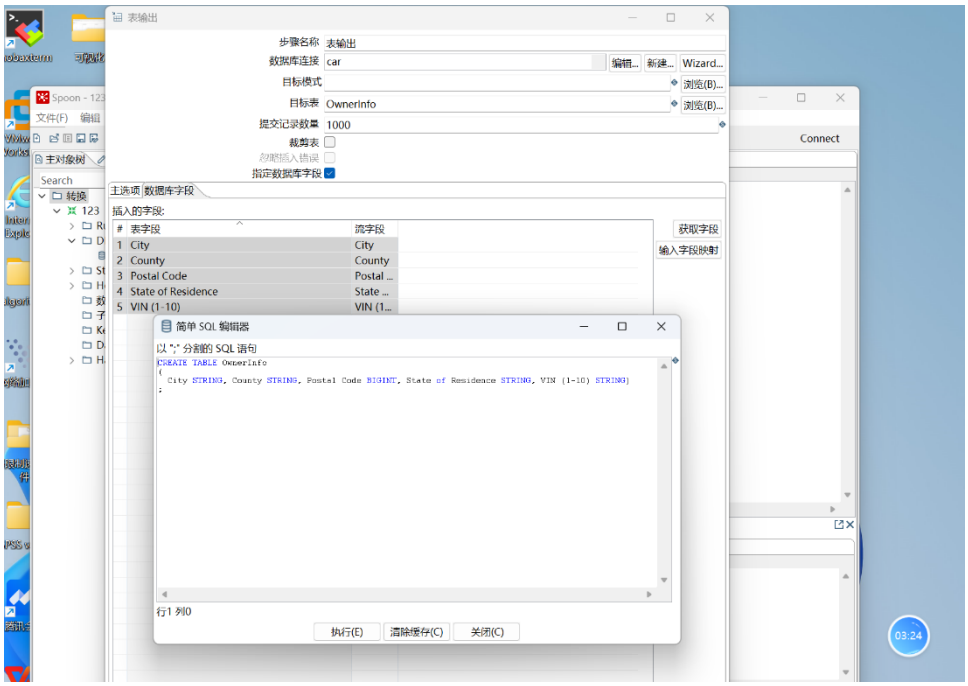
连接 hive：





新建一个转换，将我们清洗过后的 excel 文件作为输入，以表的形式输出：



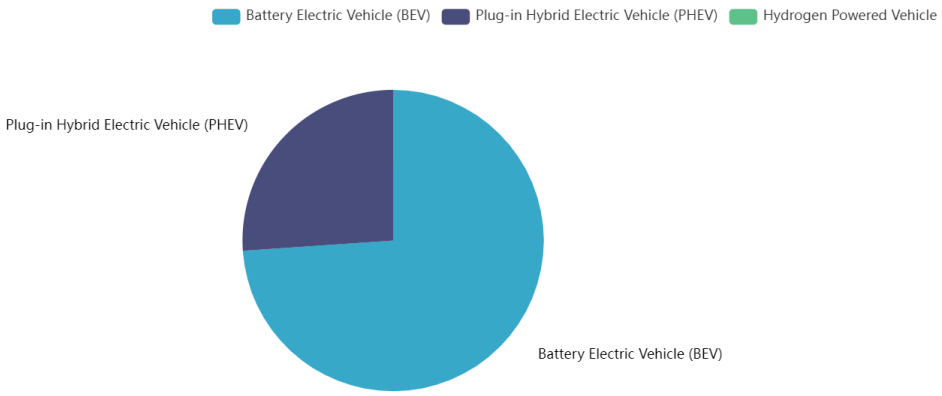


执行 SQL 语句，生成数据表。单击【SQL】按钮，弹出【简单 SQL 编辑器】对话框

六、数据可视化及数据分析

利用可视化框架 superset 进行可视化分析

1. 新能源各类车型占比



RESULTS SAMPLES

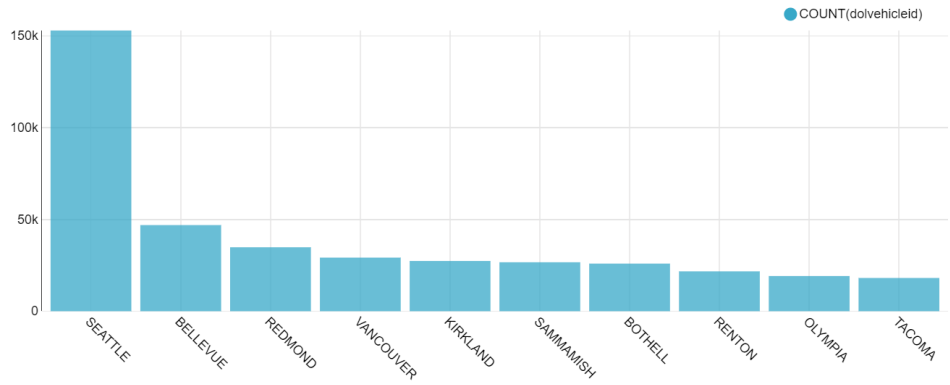
Search 3 rows

cleanalternativefuelvehicletype	COUNT(vin)
Battery Electric Vehicle (BEV)	632654
Plug-in Hybrid Electric Vehicle (PHEV)	223260
Hydrogen Powered Vehicle	3

- 市场主导趋势：从占比数据可以看出，Battery Electric Vehicle (BEV) 在新能源车型市场中占据主导地位，其占比远远超过 Plug-in Hybrid Electric Vehicle (PHEV) 和 Hydrogen Powered Vehicle。这表明消费者更倾向于购买纯电动车辆，可能是因为其零排放和更低的运营成本。
- 潜在增长机会：尽管 Battery Electric Vehicle (BEV) 占据市场主导地位，Plug-in Hybrid Electric Vehicle (PHEV) 仍然有一定的市场份额。这表明消费者对同时拥有电动和燃油驱动系统的混合动力车型也存在需求。汽车制造商可以利用这一机会继续开发和推广 PHEV 车型，以满足消费者对长途行驶和充电基础设施有限的关切，同时提供更环保的驾驶选择。
- 氢能源车型的挑战：Hydrogen Powered Vehicle 在市场中的占比非常小。这可能反映出目前氢能源车型在消费者中的认知度和接受度较低，以及相关基础设施建设和成本问题。

这些数据表明 Battery Electric Vehicle (BEV) 是新能源车型市场的主导力量，而 Plug-in Hybrid Electric Vehicle (PHEV) 仍然具有一定的市场份额。对于汽车制造商来说，抓住纯电动和混合动力车型的市场机会是至关重要的，同时需要关注氢能源车型的发展潜力和相关挑战。这些信息可以指导企业在产品开发、市场推广和战略规划方面做出明智的决策。

2. 新能源汽车购买量前十城市



RESULTS SAMPLES

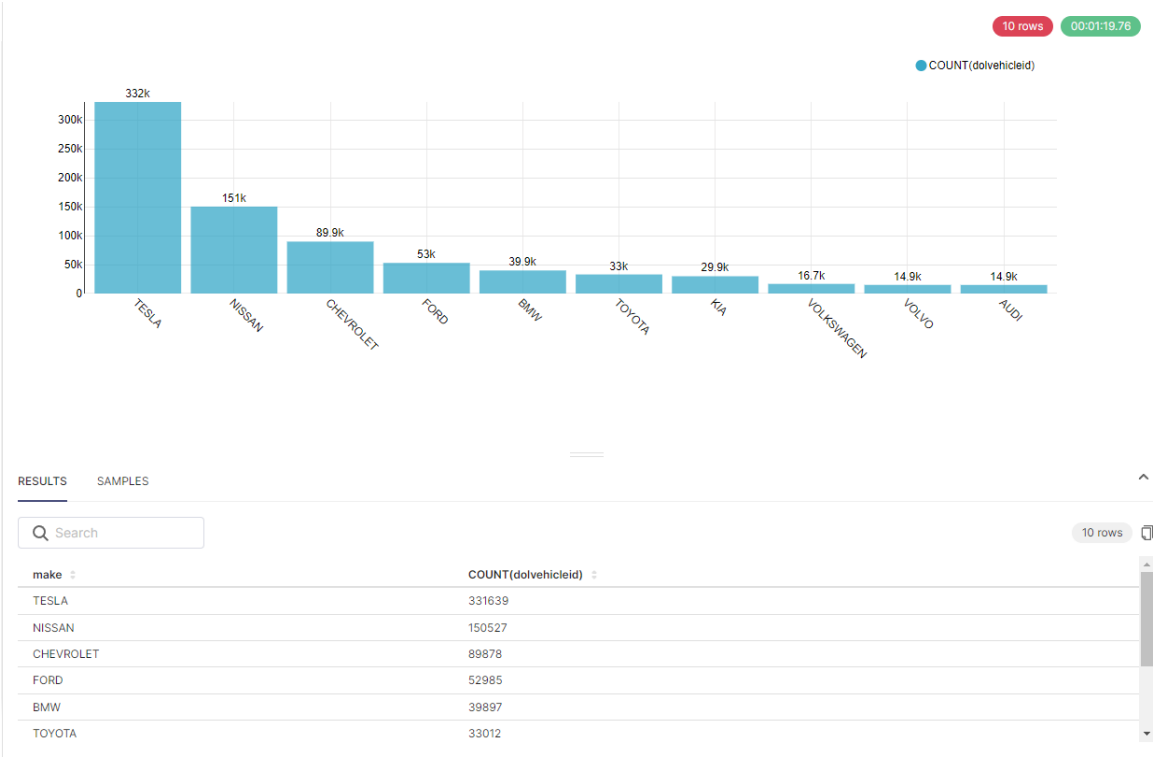
Search 10 rows

city	COUNT(dolvehicleid)
SEATTLE	152953
BELLEVUE	46933
REDMOND	34850
VANCOUVER	29220
KIRKLAND	27380
SAMMAMISH	26680

- 市场热点区域：SEATTLE 是新能源汽车购买量最高的城市，具有巨大的市场潜力。该地区可能有一系列因素，如政府支持政策、充电基础设施建设和消费者环保意识的提高，使得新能源汽车在该地区受到欢迎。
- 城市特定需求：每个城市的购买量都有所不同，这可能与城市特定的需求和特点有关。例如，除 SEATTLE 外，其他华盛顿州城市的购买量相差不大，这可能与该地区的经济状况、居民收入水平以及对环境可持续性的关注有关。
- 市场潜力分析：根据购买量前十的城市数据，汽车制造商和销售商可以进一步分析这些城市的特点和因素，以制定适应当地市场的销售策略和营销活动。他们可以关注这些城市的消费者需求、竞争对手的表现以及市场发展趋势，以更好地满足当地消费者的需求。
- 充电基础设施规划：购买量较高的城市需要更多的充电基础设施来支持新能源汽车的发

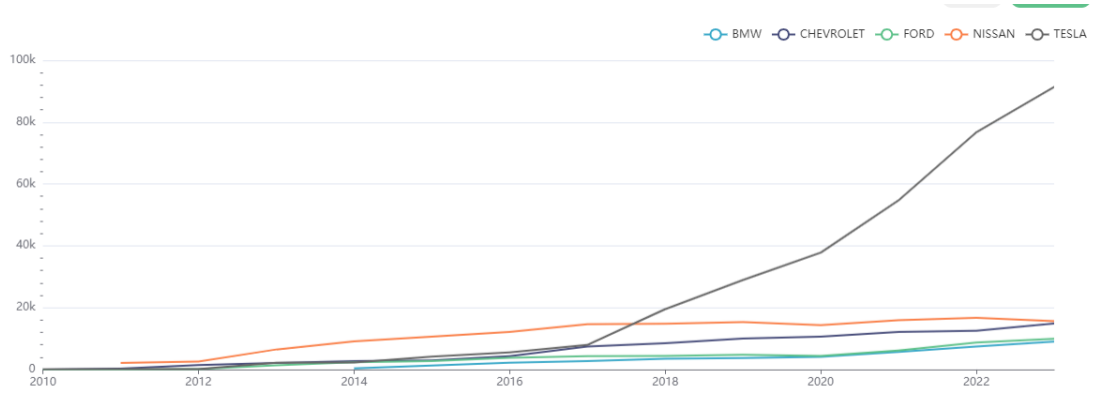
展。了解这些城市的购买量可以帮助政府和能源公司确定充电站的建设和布局计划，以满足日益增长的需求。

3. 新能源汽车销售前十厂商



- 市场领导者：TESLA 是新能源汽车销售量最高的厂商，具有明显的市场领导地位。TESLA 的销售量远远超过其他竞争对手，这可能归因于其独特的品牌形象、技术创新和市场知名度。
- 品牌竞争：NISSAN、CHEVROLET、FORD、BMW 和 TOYOTA 都在新能源汽车市场中占据一定的市场份额。这表明这些传统汽车制造商正在积极进军新能源汽车领域，并与 TESLA 等新兴厂商展开激烈的竞争。这些厂商可以继续加大新能源汽车产品线的研发和推广力度，以提高市场份额。
- 品牌认知度：TESLA 在新能源汽车市场中的销售量远超其他厂商，这可能意味着 TESLA 的品牌认知度更高，消费者对其产品更加熟悉和信任。其他厂商可以借鉴 TESLA 的营销策略和品牌建设经验，提升自身品牌的认知度和市场竞争力。
- 市场多样性：新能源汽车市场存在多个竞争厂商，消费者有更多的选择。这对消费者来说是好消息，他们可以根据自己的需求和喜好选择适合的品牌和型号。同时，这也促使厂商加大产品创新和品质提升的努力，以满足不断增长的市场需求。

4. 销量前五的品牌销售趋势图



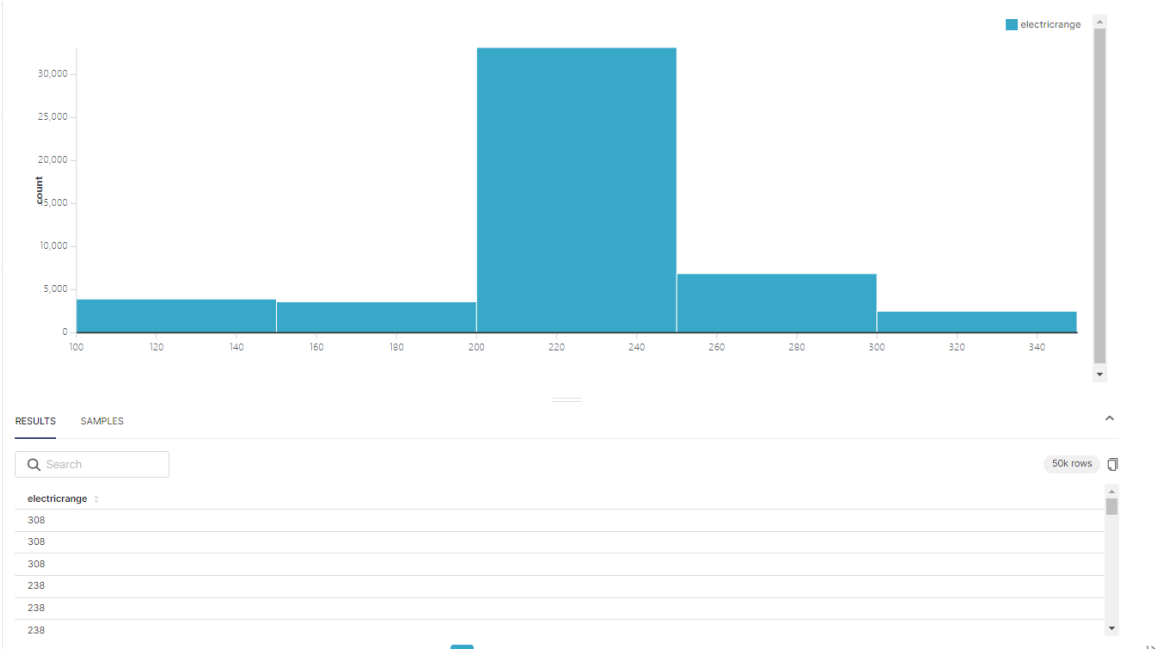
RESULTS SAMPLES

Search 14 rows

__timestamp	BMW	CHEVROLET	FORD	NISSAN	TESLA
2018-01-01 00:00:00	3485	8492	4369	14832	19516
2019-01-01 00:00:00	3692	10037	4740	15338	28951
2020-01-01 00:00:00	4092	10587	4373	14343	37789
2021-01-01 00:00:00	5675	12143	6165	15964	54721
2022-01-01 00:00:00	7449	12546	8709	16681	76719
2023-01-01 00:00:00	9010	14886	9967	15658	91399

- 销售趋势：通过观察每个品牌的销售趋势，可以发现 TESLA 的销售量呈现显著的增长趋势。这表明 TESLA 品牌在这段时间内取得了显著的销售增长，可能受到市场需求的推动和品牌的成功推广。其他品牌如 BMW、CHEVROLET、FORD 和 NISSAN 的销售量变化相对较小，没有像 TESLA 一样的明显增长趋势。
- 市场份额：通过比较不同品牌的销售量，可以获得有关市场份额的信息。在这段时间内，TESLA 的销售量明显超过其他品牌，表明 TESLA 在市场上占据了更大的份额。BMW、CHEVROLET、FORD 和 NISSAN 的销售量相对较低，可以进一步分析市场需求、竞争对手的表现以及品牌策略等因素，以提高市场份额。
- 品牌竞争力：通过比较不同品牌的销售量，可以获得关于品牌竞争力的信息。TESLA 作为销量最高的品牌，显示出强大的品牌竞争力和市场需求。其他品牌如 BMW、CHEVROLET、FORD 和 NISSAN 则需要进一步分析市场表现、产品定位和品牌推广等因素，以提高自身的竞争力。
- 市场趋势：通过观察不同品牌的销售趋势，可以获得有关整体市场趋势的信息。TESLA 的销售量持续增长可能反映了新能源汽车市场的快速增长和消费者对电动汽车的日益认可。这可以为其他品牌提供有关市场发展趋势和消费者偏好的信息，以制定相应的战略和产品规划。

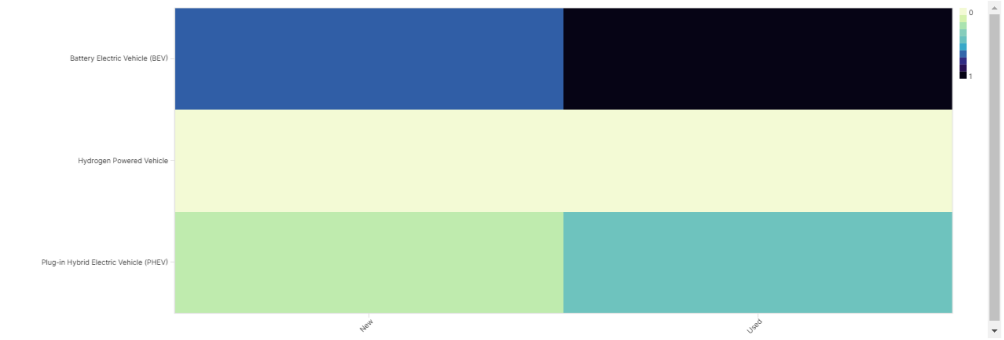
5. 续航里程程分布直方图



- 用户需求：续航里程集中在 200 到 250 的范围内可能表明中等续航里程的电动汽车在市场上具有较高的需求。这可以为汽车制造商提供有关消费者对电动汽车续航里程的期望和偏好的信息。制造商可以根据这一需求开发和推出具有适当续航里程的产品来满足市场需求。
- 市场机会：如果续航里程集中在 200 到 250 的范围内，而市场上尚未有充分满足这一范围的竞争对手产品，那么这可能是一个市场机会。制造商可以考虑开发续航里程在这一范围内的电动汽车，以填补市场空缺并满足消费者需求。

我们可以看出在 100-200 续航里程范围，通过增加成本、优化技术来提高里程所获得的需求增幅大、性价比高，并且 200-250 续航历程范围已经可以满足大部分用户的需求，250 以上里程范围再提高续航里程，性价比较低。

6. 不同清洁能源车型与车辆状态的相关关系



RESULTS SAMPLES

Search 6 rows

neworusedvehicle	cleanalternativefuelvehicletype	COUNT(dolvehicleid)
New	Battery Electric Vehicle (BEV)	256699
New	Hydrogen Powered Vehicle	2
Used	Plug-in Hybrid Electric Vehicle (PHEV)	160365
Used	Battery Electric Vehicle (BEV)	375718
Used	Hydrogen Powered Vehicle	1
New	Plug-in Hybrid Electric Vehicle (PHEV)	62888

- 电动车型(Battery Electric Vehicle - BEV)在新车和二手车市场上都有很高的销量。这表明消费者对电动车的需求较高，无论是购买新车还是二手车。因此，针对电动车的销售和服务可以是一个有前景的商业领域。
- 氢燃料车型（Hydrogen Powered Vehicle）在新车市场上的销量相对较低，而在二手车市场上销量几乎可以忽略不计。这可能表明目前市场对氢燃料车型的需求较低，商业机会有限。在投资和发展方面，可能需要更多的市场研究和推广工作。
- 插电式混合动力车型（Plug-in Hybrid Electric Vehicle - PHEV）在二手车市场上有一定的销量，但在新车市场上的销量相对较低。这可能意味着消费者对插电式混合动力车型的兴趣在一定程度上集中在二手车市场。商家可以考虑提供更多的二手插电式混合动力车型以满足市场需求。
- 不同清洁能源车型在二手车市场上的销量普遍较高。这可能是因为二手车市场提供了价格更低的选择，吸引了更多消费者购买清洁能源车型。商家可以考虑加大对二手清洁能源车型的销售和推广力度。

七、总结

通过集成 Hadoop、Hive 和 Presto 等大数据技术，我们创建了一个强大的数据仓库，用于存储和分析华盛顿州电动汽车车辆登记与所有权变更活动数据集。在我们的总体设计中，Hadoop 作为分布式存储和计算框架，为大规模数据提供高可靠性和可扩展性。Hive 作为元数据存储和查询引擎，通过 HQL 转换为 MapReduce 任务，支持复杂的 SQL-like 查询。Presto 则作为交互式查询引擎，连接 Hive 等数据源，实现低延迟的数据分析。总体而言，我们完成了一个完整的数仓的 workflow，最后取得了很多可视化的结论分析并挖掘出了很多有用的信息，小组的每个人都收获良多。