大数据分析案例：
Hadoop 在铁路客票分析领域的应用

需求背景：

随着高铁建设的进程不断推进，为了科学合理的设计铁路开行方案，提高整体运营效率，需要对铁路客流进行分析，分析铁路流量的变化规律，预测铁路铁路客运量，为精细化的客票营销提供科学的数据支撑。以运输理论和客运信息资源为基础，运用信息技术、运筹学和数理统计分析等方法为客运部门把握市场动态、预测客流量的变化趋势，制定列车开行方案、提供辅助决策的依据、数据管理的功能，该系统能够协助路局客运处，使客流预测过程更科学，整个客运营销过程具有可追溯性，预测准确性评判具有可量化性。

为什么是 Hadoop

Hadoop 是 Apache 组织的一个开源的分布式计算框架，框架中最核心的设计是 HDFS 和 Map-Reduce 。基于二者 Hadoop 有一系列子项目，它们广泛用于海量数据处理，非结构化数据存等领域。

我们结合季节，时间，区域，节假日等诸多因素的影响，利用上海铁路局 2008-2012 四年（超过 10T ）的数据进行了分析。对于 T 级别数据，普通的单个服务器的数据库系统处理起来效率偏低，这样的问题只有使用 Hadoop 这样的分布式系统进行处理，才能满足需求。

HDFS 简介

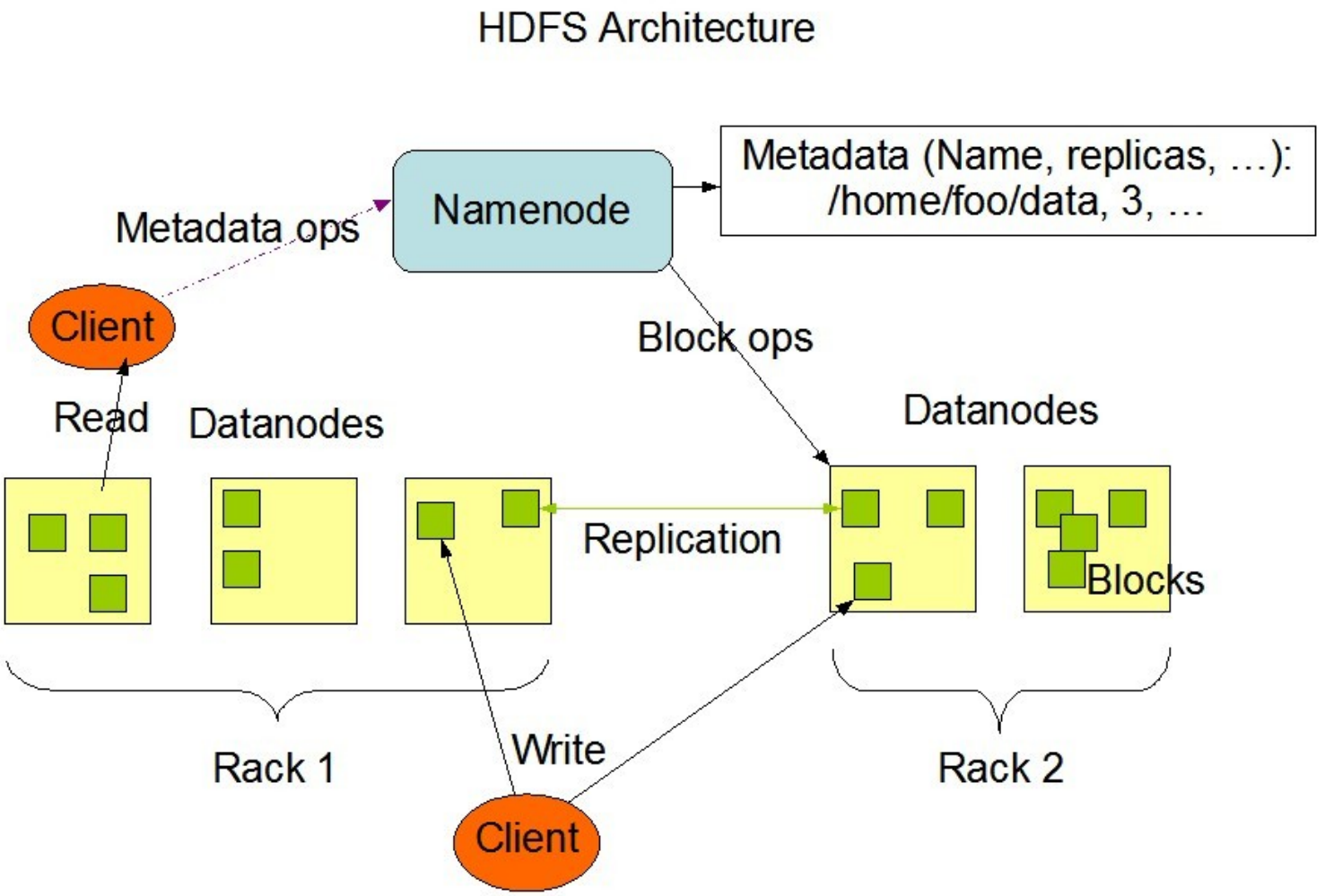
HDFS (Hadoop Distributed FilSystem) Hadoop 提供的基础设施，Hadoop 其它子项目均依赖于 HDFS 作为一个分布式文件系统，HDFS 用于部署在低成本的硬件之上（使用普通 PC 的硬盘），因此具有很高的容错性。

HDFS 是根据 google 的 GFS 实现的是开源版的 GFS 文件系统，其基本思想来源于 google 的

Bigtable: A Distributed Storage System for Structured Data

这篇论文。

HDFS 示意图


HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

MapReduce

Map-Reduce 是一个分布式的计算框架，用于大规模数据集的并行运算。 Map-Reduce 大大降低了分布式计算的难度。

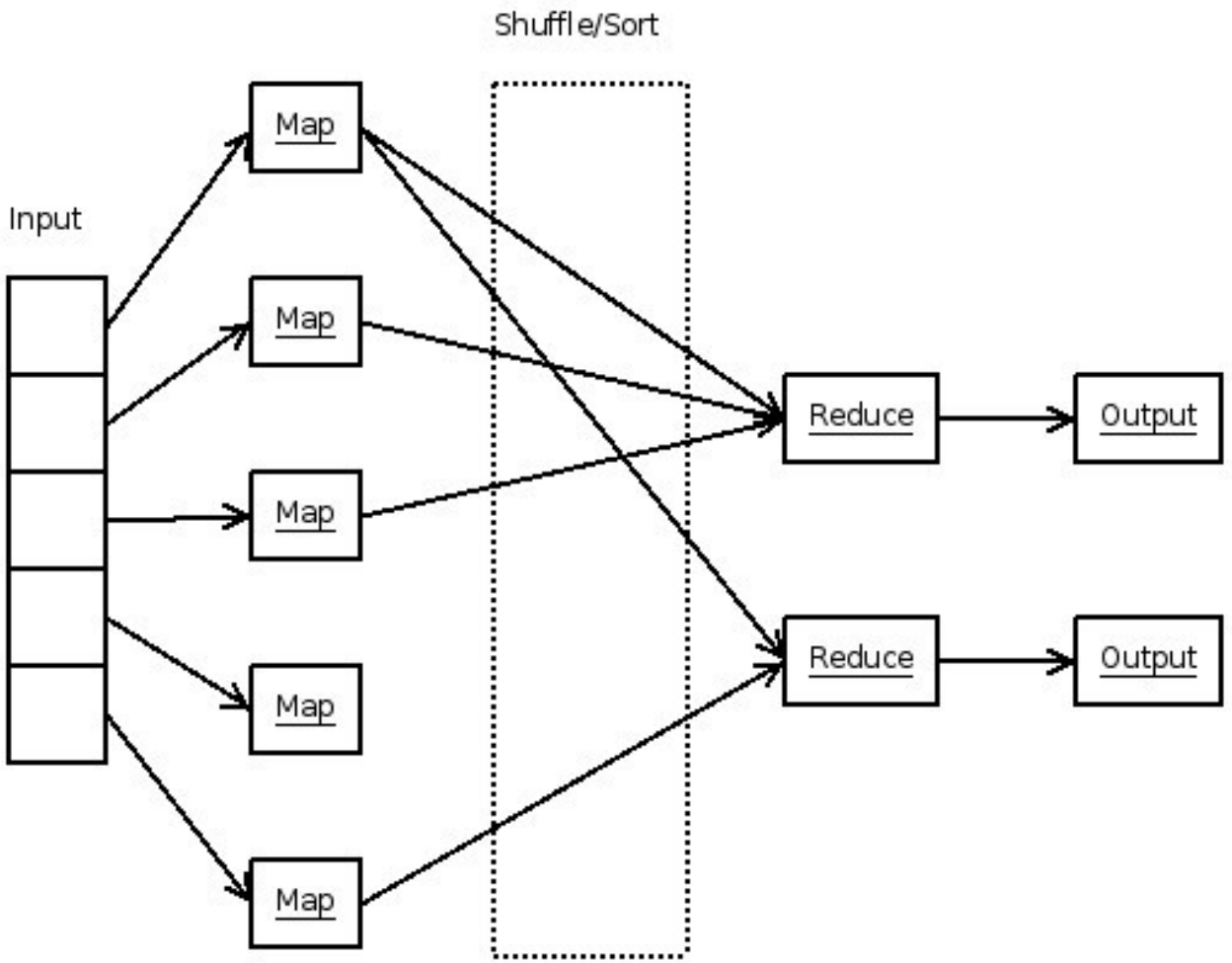
一个 Map/Reduce 作业的输入和输出类型如下所示：

(input) <k1, v1> -> map -> <k2, v2> ->

Combine →

<k2, v2> -> reduce -> <k3, v3> (output)

Map-Reduce 示意图



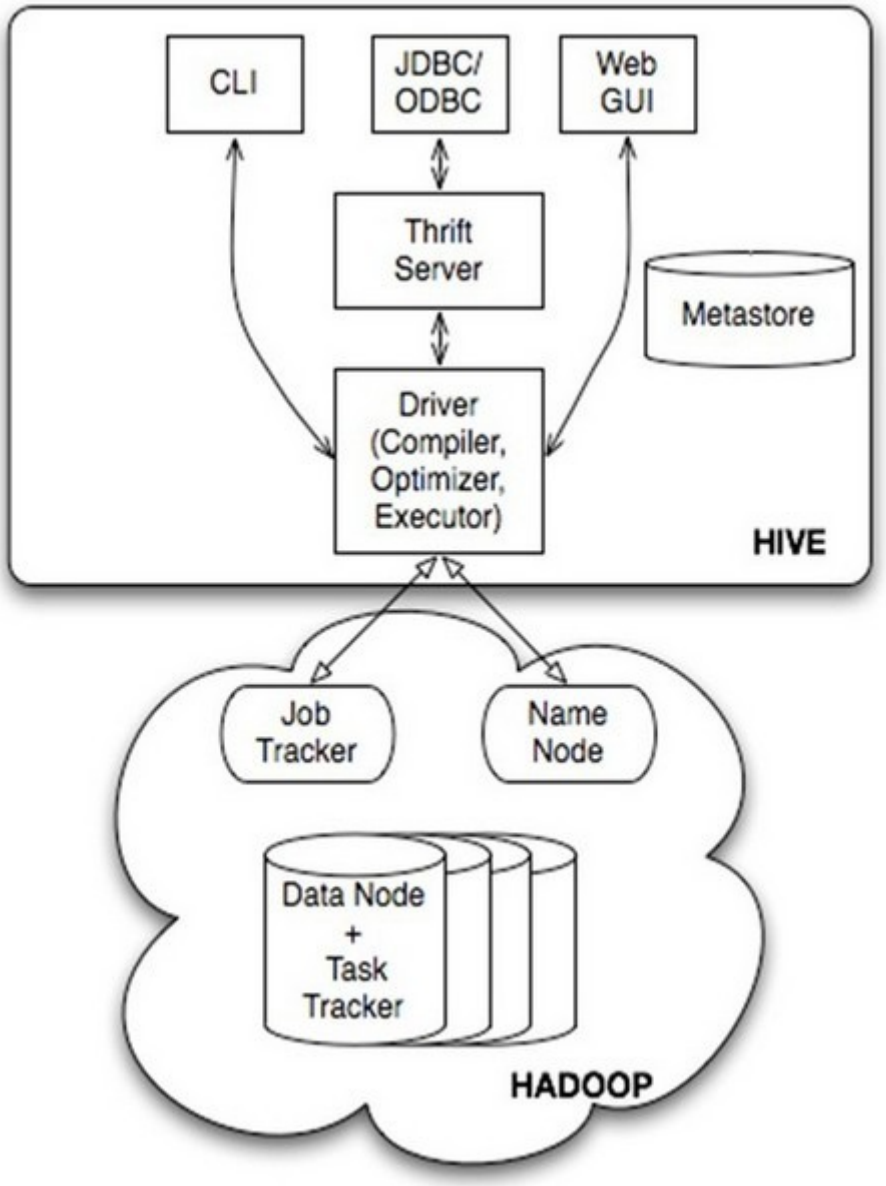
Hive 简介

Hive 是一个基于 Hadoop 的数据仓库分析框架。Hive 定义了一种类似于 SQL 的语言 -HQL，使用 HQL 可以方便使用 SQL 类似的语句分析数据，大大降低了数据分析的难度。

Hive 支持 Map-Reduce 。Hive 支 UDAF（User Defined Aggregate Function，用户自定义的聚合函数）。

Hive 是此项目中使用最多的工具，使用 Hive 的使用大大提高了我们的工作效率，尤其对去熟悉 SQL，但是并不怎么擅长 Java 程序编写的数据分析师 .

Hive 的架构


CLI
JDBC/
ODBC
Web
GUI
Thrift
Server
Metastore
Driver
(Compiler,
Optimizer,
Executor)
HIVE
Job
Tracker
Name
Node
Data Node
+
Task
Tracker
HADOOP

客票数据分析平台系统简介

一、硬件

1. 集群规模：总共 30 个节点，其中 NameNode 一个，Datanode28 个。

2. 每个节点的机器配置：

a. 硬盘 300G

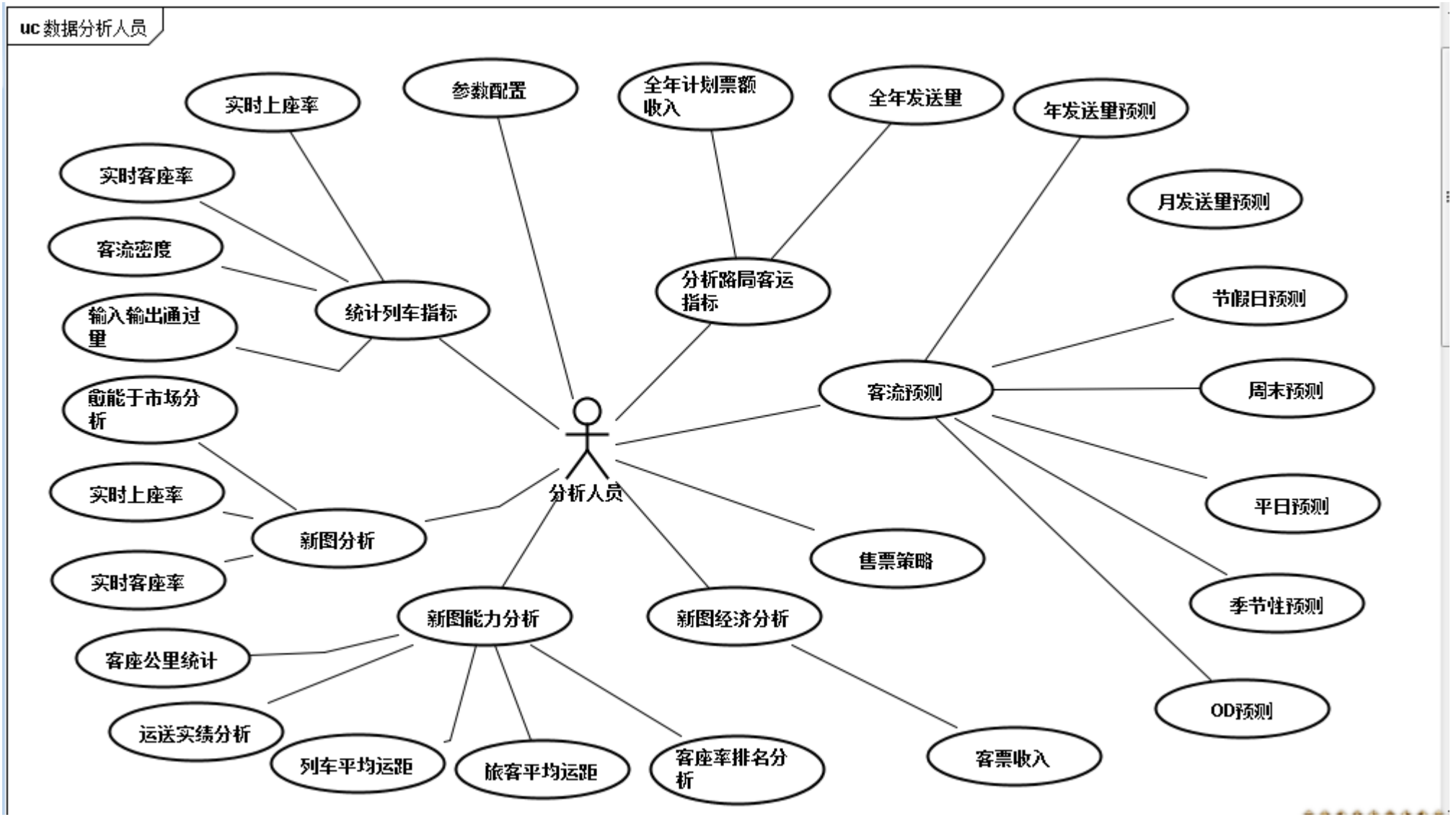
b. 内存 20G

3. 集群网络：千兆网络

二、软件

1.Hadoop-1.0.4 Hive-0.9.0 Hbase-0.94

业务结构与数据流


uc 数据分析人员
实时上座率
参数配置
全年计划票额收入
全年发送量
年发送里预测
实时客座率
月发送里预测
客流密度
分析路局客运指标
统计列车指标
节假日预测
输入输出通过里
客流预测
周末预测
愈能于市场分析
分析人员
实时上座率
平日预测
新图分析
售票策略
实时客座率
季节性预测
新图能力分析
新图经济分析
客座公里统计
OD预测
运送实绩分析
客票收入
列车平均运距
旅客平均运距
客座率排名分析

业务范围

由于客流预测目前处于探索阶段，一期客流预测为降低业务复杂度，将预测范围局限于上海局管内客运专线高铁线路。根据建设目标内容，考虑到客流的稳定性，挑选具有一定业务代表性的线路进行研究，将预测对象设定为沪宁高铁全线。根据调研成果，路局客运处关心的预测重点为节假日预测，因此一期客流预测的预测范围只限于节假日期间内沪宁线 OD 区间的客流预测。

1. 铁路运输企业历年运输总量分析
2. 铁路运输企业年月客流预测
3. 节假日客流预测
4. 平日客流预测

业务范围

5. 周末客流预测

6. 始发站客流预测

7. 季节性客流预测

8. 统计车次客座率

9. 统计线别客流密度

10. 净输入输出通过客流统计

.....

基础数据（票务存根）

我们利用一个路局的既有信息系统中 2008-2012 四年的所有票务存根，票务存根中包含四年中所有本局发送旅客的车票信息，所有的数据都是 .bcp.gz 格式的数据库压缩文件，通过编写 shell 脚本将 2T 大小的原始数据解压转化格式和编码，并自动发送到指定的位置，由于数据量较大，此项工作的脚本程序大概需要 4 个小时才执行完。

利用 Hive 构建数据仓库

在搭建好的 Hive 环境中创建表，类似于关系型数据库中创建表的方法，然后将存储在 HDFS 中的数据载入到 Hive 环境中，载入的速度非常快，因为对于 HDFS 来说，数据没有移动只是元数据结构发生了变化，由于数据量太多，我们采用编写 shell 脚本自动生成 HiveQL 文件然后使用

```
hive -f LoadSsript.sql
```

提交任务。

利用 RHive 分析数据

在 hive 中我们可以像操作数据库那样写查询脚本，从海量数据中抓取我们需要的数据，然后将抓取的数据放在 R 环境中进行建模分析。

- R 是一个非常方便的数据分析环境，通过使用 RHive 这个包通过 hiveserver 服务访问 Hive 中的数据。
- 将 Hive 快速查询海量数据的功能和 R 快速实现各种数据分析方法的功能结合起来是大数据分析的一个不错的选择。