汽车行业大数据分析架构设计及实施方案研究

信息管理与信息系统 13　方佳超

指导教师　李艳

摘要

处于大数据时代，海量数据无时无刻不在生成，其间蕴含着无限商机，影响着各行各业。一些传统企业在越来越多的创新企业面前，危机感不断加强，为了提高自身的竞争优势，企业转型势在必行，是一个不二的选择。于是，这些传统企业也加入了该行列。每个行业由于自身业务的独特性以及数据的特性，市场上通用的大数据架构方案并不足够，更需要的是适用于该行业的数据架构及解决方案。

本文以汽车行业为研究对象，主要目的是为汽车行业设计大数据分析平台的基础架构。本文首先阐述了国内外大数据的研究现状，并对课题提到的相关技术做出简要的介绍。然后针对汽车行业的数据及业务特殊性做了需求分析，并设计了汽车行业大数据分析平台的基础架构。最后，以奥迪中国为例，实际应用了本文所提出的大数据分析基础架构，并为其提供了实施管理方案。

关键词：大数据，大数据架构，数据分析，需求分析，汽车行业

Design and Implementation of Big Data Analytical Architecture for Automotive Industry

Information Management and Information System 13 Fang Jiachao

Supervisor Li Yan

Abstract

Living in a big data era, massive data generates all the time, which brings plenty of business opportunities and influences all walks of life. Facing more and more innovative start-ups, traditional industries have started to feel threatened. In order to improve their own competitiveness, enterprise transformation is crucial. Therefore, these traditional industries seize the opportunities and face the challenge of big data. Because of the uniqueness of business and data of each industry, general big data solution in the market is not enough, which means that customized architecture solution for each industry is required.

This thesis uses automotive industry as a research object, in order to design a basic data analytical infrastructure for automotive industry. This thesis summarizes the current research of big data at home and abroad and introduces the related big data technologies. Then, the author conducts a requirement analysis and designs basic big data architecture for automotive industry. At last, the thesis applies the designed architecture and provides a customized management solution to Audi (China).

Key words: Big data, big data architecture, data analysis, requirement analysis, automotive industry

目录

图表目录

1 绪论

1.1 研究背景

有一个耳熟能详的统计结论，世界上，90%的数据是过去 5 年产生的。实际上，事实可能更让人吃惊，在过去三十年里，全世界的数据量大约每两年增长 10 倍，这一数字远超计算机领域的摩尔定律。

毋庸置疑，当今社会拥有的数据量，比历史上任何时候都多得多。多样的数据来源也为数据量的增长带来了更多可能。这些数据来源多种多样，有收集环境信息的传感器、图片视频、社交网络上的信息、位置信号等等。据 IDC 发布的数字宇宙研究报告，至 2020 年，预计世界上总数据量将超过 40ZB，这意味着，每个地球人将产生 5200GB 的数据[1]。

大数据时代已经到来，日后各行各业的商业决策将更加地取决于数据分析的结果，而非以往的基于经验与直觉。仅仅五年，数据革命牵动着各行各业，大数据时代的来临，不仅变更了人们的思维方式，更引领了诸多传统企业的变革。2015 年以来，大数据解决方案在全球范围内不断成熟，大数据应用在各领域全面展开，推动大数据技术的强劲增长的同时，也带动了各个领域的发展。

2016 年，全球大数据市场整体增长 22%，达到 453 亿美元，而这一数字，预计在 5 年之后，2022 年，将达到 802 亿美元之多。而大数据技术对全球 IT 支出的间接推动仅 2016 年一年就超过 2300 亿美元[2]。行业变革初现端倪，大数据的应用最初集中在物流管理、医疗健康以及金融投资等方面，逐渐蔓延至各行各业，更多的传统企业在其中看到了商机，加入到了这个行业中来。

传统汽车企业也不例外，这些企业的核心业务越来越离不开车联网。海量的车辆数据时刻在生成，让这些传统车企看到了潜在的业务价值，纷纷加入大数据应用的行列，甚至开始思考建设自身的大数据平台。然而，企业转型没那么简单，市场上没有汽车行业通用的大数据解决方案，主流方案实施困难而且成本太高，让很多传统车企望而却步。

1.2 研究目的及意义

本文主要的研究目的是设计出在汽车行业通用的大数据分析基础架构，为汽车行业提供更适合自身的大数据解决方案。

由于每个行业业务的独特性以及数据的特点，通用的企业大数据架构并不能满足每个行业各自的需求。越来越多的学者开始针对不同的行业以及领域，提出相应的大数据架构方案，如：电力企业、电信运营商、石化企业等。但是，在文献资料中，均没有查阅到针对汽车行业设计的大数据架构。而大数据是时代的趋势，目前，大众、宝马、奔驰等传统汽车企业都准备开展或者已经开展自身的大数据业务，所以汽车行业未来在大数据方面的应用前景十分可观。由于车联网的流行以及车载传感器技术的进步，不同于其他行业，汽车行业的数据很大一部分是通过传感器收集的机器数据，因此，更有

必要单独研究汽车行业的大数据基础架构。这样，不仅能够为车企建立自身的大数据平台提供参考，同时也能够为大规模机器数据的分析提供借鉴意义。

1.3 研究方法

大数据平台的基础架构分为硬件系统以及软件系统，而本文提出的大数据基础架构侧重于软件系统。所以，本文中，将汽车行业大数据平台看作一个软件系统，对其基础架构的设计，遵循软件工程中软件设计的基础原则，选择目前最流行的企业大数据架构方法，大数据业务需求出发，分析所需的数据，根据数据源收集、存储以及整理的要求，总结出汽车行业大数据平台的功能性需求与非功能性需求，再根据这些需求设计汽车行业大数据分析的架构方案，最后通过奥迪中国的大数据平台作为实际案例，具体实施该架构方案，证明了该大数据架构方案的可行性。

1.4 论文的主要研究内容

本文选题基于大数据时代，传统汽车行业变革的需求以及奥迪中国企业管理有限公司的实际工作需要，根据本人在奥迪中国的实际工作和项目经历，在导师的指导下进行研究。本文的研究目标是为汽车行业设计通用的大数据分析的基础架构，并应用该为奥迪中国提供大数据分析平台的基础设施建设及实施管理方案。

本文研究内容一共分为七个部分，对大数据背景下的理论和方法在企业数据分析平台的应用进行论述：

第一部分，绪论。讲述了研究背景，阐述了课题的由来以及研究意义；介绍了该课题的研究方法与内容，并理清了论文的组织结构。

第二部分，文献综述。阐述了国内外大数据研究现状，总结了三大典型的大数据应用，分析了流行的企业大数据解决方案以及常见的大数据架构方法，最后介绍了大数据相关理论与技术。

第三部分，对汽车行业大数据分析架构的设计进行了需求分析，总结了功能性需求与非功能需求。

第四部分，根据第三部分需求分析的结果，根据信息的价值链流程，设计了汽车行业大数据分析架构，并具体介绍了数据收集、数据存储以及数据分析的各个组件的选择原因，最后论证了架构能够满足第三部分中所提出的需求。

第五部分，对奥迪中国现有的数据基础架构进行分析，结合企业提出的需求，应用第四部分所提出的大数据架构方案，为奥迪中国设计大数据基础架构，验证了该架构方案的可行性。

第六部分，设计奥迪中国数据分析平台的实施管理方案，主要针对在该平台上运作的数据分析项目设计了组织管理以及业务流程。

第七部分，总结与展望，对汽车行业大数据分析基础架构的设计及针对奥迪中国的应用进行了总结，并展望了未来的改进方向。

1.5 论文的组织结构

本文的组织结构如下图 1.1 所示。

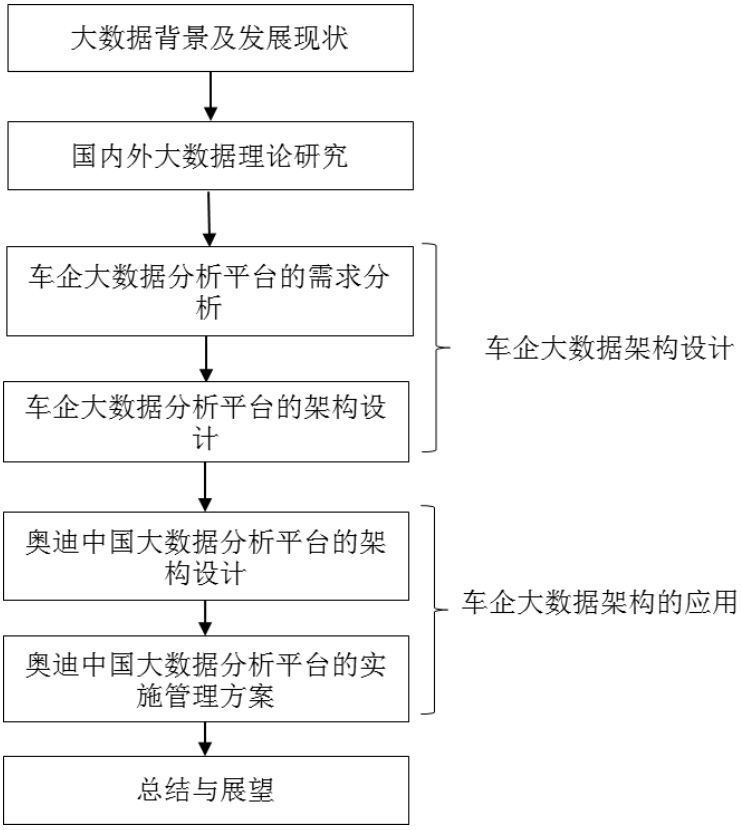

图 1.1 论文的组织结构

Fig.1.1 The architecture of the thesis

2 相关研究综述

2.1 国内外大数据研究现状

2003-2006 年间，Google 公司先后发表关于 GFS[3]、MapReduce[4]、BigTable[5]等与大数据技术相关的论文，引起了诸多互联网公司，诸如 Facebook、Yahoo 的关注，这不仅是大数据发展的起源，也是目前应用最为广泛的大数据处理开源平台 Apache Hadoop 的诞生伊始。

2009 年，联合国开始“全球脉动计划”，其目的便是利用大数据的应用带动落后区域发展；2012 年，世界经济论坛年会，“大数据，大影响”也成为最重要主题之一。2009 年开始，美国政府便公开了 40 万条属于政府数据库的原始数据记录，并将“大数据战略”纳入国家安全、创新及网络信息安全战略。同年，美国政府提出《大数据研究和发展倡议》，并投资了 2 亿美元用于支持大数据核心技术的研究与应用。紧跟趋势，英国政府也将大数据技术作为发展重点，投资了 6 亿英镑在 8 个高新技术中，其中，投资在大数据中的比例占 30%。2014 年，欧盟呼吁各成员国积极响应大数据时代的到来，将大数据发展放在最重要的战略地位[6]。

学术方面，美国麻省理工大学计算机科学与人工智能实验室下成立了大数据科学技术中心，积极和微软、英特尔和 EMC 等公司进行合作研究。这些研究集中在数据共享、存储和大数据解决方案，尤其关注医疗、企业与行业计算以及数据加密等领域[7]。英国在这方面也紧随美国，第一个大数据医疗卫生研究中心在牛津大学成立，旨在带动医疗数据分析的发展。初次之外，在匈牙利，偶中核子中心也建立了一个超带宽数据中心，这是欧洲目前具有最强传输能力的数据处理中心[8]。

产业方面，IBM，DELL，Microsoft，HP 等国际知名企业都将大数据作为重点研究内容，提出了各自的大数据解决方案和应用。2005 年期，IBM 投资 160 亿美元收购了 35 家进行大数据分析的公司。同时，与全球近千所高校进行了大数据相关研究，行业应用及教学等等的全方位合作。

不仅国外如此，我国政府、学界及产业界也对大数据的研究十分重视，并且也都制定了相关的研究计划。

2012 年 3 月，我国科技部发布《十二五国家科技计划信息技术领域项目征集指南》，其中明确指出“面向大数据的关键技术”。地方政府对于大数据的发展也是高度重视。2013 年，上海提出了推动大数据发展的相关计划，重庆市也有大数据行动相关计划；2014 年，广东省成立了大数据管理局，负责所有大数据相关的战略规划，大大促进了大数据研究的发展[9]。

学术方面，国内的诸多高校，研究所纷纷成立了自己的大数据研究中心，还有很多相关的学术组织以及学术活动。2012 年，中国计算机学会以及中国通信学会成立了大数据专家委员会；同时，有很多相关的论坛和会议，如中国大数据技术大会，中国国际大数据大会，大数据分析以及管理国际研讨会等[10]。

产业方面，百度，腾讯，阿里巴巴等知名互联网企业都成立了各自的大数据研究实验室，并与诸多高校进行了合作与交流。

总而言之，国内外无论在政府、学术还是产业方面，都充分意识到了大数据的重要性：从政府开始实施鼓励大数据相关政策，学术界积极研究大数据相关理论与技术，业界投资大数据领域并且与学界积极进行研究合作。

2.2 大数据的典型应用

随着大数据研究的深入，大数据的应用也在不断地增多。典型的大数据应用有:企业内部大数据应用、物联网大数据应用、社交网络大数据应用等[11]。

企业内部的大数据应用经常涉及销售、运营、供应链等方面，如利用大数据来预测消费者行为。最新的企业内部大数据应用还涉及人力资源管理，利用企业内部的员工档案等数据，可以预测员工离职率等。

物联网又称传感网，是将各类传感设备与互联网连接起来的网络。传感设备主要有四类：听觉接受设备、视觉接受设备、感觉接受设备和运动接受设备。因为物联网的发展，数据量级不断增大，又为大数据应用创造了更多可能。智慧城市就是典型的物联网应用，具体举个例子就是智能交通，通过汽车 GPS 定位信息等，可以优化道路交通。

社交网络大数据应用集中在对于流行社交应用如 Facebook、Twitter、微博等的用户动态进行深度挖掘，往往会利用到数据挖掘、机器学习等技术。这方面典型的应用有用户行为预测，如：音乐推荐、智能广告、风险预测等。

综上所述，以上三类是最为典型的大数据应用，任何行业想要充分利用大数据，挖掘数据当中的价值，都离不开这三大类的应用。

2.3 企业大数据解决方案

在企业有了大数据业务需求之后，为了更好的迎接大数据带来的机会和挑战，充分发挥其带来的业务价值，就需要一个相比传统架构更加灵活，易扩展，可管理的适用于大数据分析的数据基础架构。目前，国内外流行的企业大数据解决方案大多是以 Hadoop 为核心的分布式软件系统或者软硬件一体化的解决方案。以下，简单介绍一些目前主流的企业大数据解决方案[12]。

英特尔：英特尔公司提供的大数据解决方案为英特尔 Hadoop 发行版，里面的组件包括了分布式文件系统（HDFS）、分布式计算框架（MapReduce）、分布式数据库（HBase）、数据仓库（Hive）、数据处理（Pig）还有数据挖掘工具（Mahout）。该解决方案的优点是对于在英特尔平台上运行的 Hadoop 及其组件进行了一系列优化。

EMC：EMC 提供的大数据解决方案叫做 EMC Greenplum，其主要组件有 Greenplum 数据库和 Greenplum Hadoop，并专门为大数据分析提供了 Greenplum DCA 一体机。该解决方案的优点能够高效处理结构化、非结构化和半结构化数据。

谷歌公司：BigQuery 是谷歌公司提供的大数据解决方案，其主要组件有分布式文件系统（GFS，Google File System)、分布式计算框架（MapReduce）以及分布式数据存储系统（BigTable）等。事实上，Hadoop 的理论基础就是谷歌公司的这三大组件。该解决方案的特点是，利用云计算，能够高效处理 TB 级别的大文件。

IBM：InfoSphere 是 IBM 公司提供的数据分析解决方案系列，目前有 BigInsights 和 InforSphere Streams 两个大数据分析平台，为企业提供了全面的大数据解决方案。InfoSphere 的特点是可与传统的关系型数据库进行集成，更适合企业级应用。

甲骨文：甲骨文公司为企业提供软硬件一体化的大数据解决方案，包括操作系统（Oracle Linux）、非关系数据库（NoSQL DataBase）、应用开发包（Oracle JDK）等。该解决方案的特点是软硬件一体化，易于实施。

微软：微软有一系列支持大数据的产品与服务，包括在 Windows 平台上提供基于云的 Hadoop 服务、集成 Hadoop 版本的 Windows 服务器和数据分析平台（Microsoft BI）。该解决方案的特点是支持企业用户在原有的微软软件的基础上快速部署大数据解决方案。

目前，国内外没有专门为汽车行业大数据分析设计架构的文献。而汽车行业由于其数据以及业务的特点，开展大数据业务能为汽车企业带来巨大的利润，所以很多传统的汽车企业开始考虑搭建自身的大数据平台。但是由于没有专门针对汽车行业大数据业务设计的架构，这些企业往往就只能选择通用的解决方案。这些解决方案有如下两大缺点：

第一，软硬件一体化的解决方案往往需要企业的 IT 基础架构进行大规模的变化，企业原有的硬件很多可能都不能继续使用，会增加企业的成本；

第二，这些通用的大数据解决方案虽然功能很全，包括各种大数据工具，但是很多实际上是企业在日常业务中用不到的，这样既对企业实施大数据方案造成困难同时也会增加企业成本。

2.4 企业大数据架构方法

目前，对于企业大数据的架构设计，尚未有学者提出一套完善的理论体系。市场上，各大提供解决方案的供应商所用的方式也不尽相同。上文中已经提到过，企业大数据解决方案分为软硬件一体化以及以 Hadoop 为核心的分布式软件系统两大类。依据这个标准，企业大数据架构方法也可大致分为两大类。第一类，软硬件一体化的大数据解决方案通常采用分层架构的方法；第二类分布式软件系统类型，通常采用软件工程中，软件系统架构的方式，这也是本文使用的方法。以下，具体介绍第二类软件架构的方法。

软件架构设计的一般流程是：业务分析、解决方案设计、系统功能设计、系统架构设计、技术体系设计[13]。大数据分析架构的设计流程以该流程为基础，分为以下几步：

第一，根据业务问题判断出其大数据类型。大数据类型主要有，机器数据、网络数据和传统数据三大类；

第二，根据大数据类型对其数据的特点进行分类。按照大数据类型对业务问题划分，能够更容易地看出这些大数据的特点。利用这些特点，我们可以更好地选择大数据收集、存储以及处理的工具[14]；

第三，开始系统功能设计。根据第二步中，总结的大数据特点，分析出这些数据的收集、存储及处理分析需求之后，就可以根据这些需求进行系统功能的设计；

第四，系统架构设计。大数据架构有两种类型，价值链架构和层次架构。由于本方案之前的步骤都是按照信息的价值链来分析设计的，所以使用价值链架构的方式更加合适；

第五，技术体系架构。这一步是针对系统接口、数据存储等具体实现进行技术规划。

2.5 大数据架构设计相关理论与技术

根据价值链架构理论，大数据架构的设计也应该从数据收集、数据存储、数据处理和数据分析四个方面来讨论。以下具体介绍在本文提出的汽车行业大数据分析架构设计过程中涉及到的大数据相关的理论与技术。

2.5.1 数据收集

大数据的数据源很多，包括了各类结构化数据、非结构化数据和半结构化数据[15] 。适用于海量数据的采集方式，主要有以下三种：日志采集、网络数据采集和数据库采集。

企业的信息系统时时刻刻在产生大量的日志数据。而日志采集平台可以收集大量的日志数据提供给数据分析平台进行数据分析。日志采集平台具体高可靠性、可扩展性和可用性的特点。常用的海量日志收集工具有 Flume 和 Scribe。

社交网络数据一般通过网络爬虫或者通过调用开源 API 的方式进行采集。通过网络爬虫或者开源 API，能够抓取网络中的非结构化信息如图像、音频还有视频等，并将其转化为结构化数据存储下来。

数据库采集是最为传统的数据采集方式。MySQL、Oracle 都是传统的关系式数据库，目前多数企业仍利用这些数据库收集企业的日常运营数据。

2.5.2 数据存储

在数据的爆发式增长、应用规模扩大以及用户高并发访问等原因的影响下，传统的数据存储方式不再能满足大数据环境下对于数据存储的需求；最适合大数据存储的就是分布式文件系统、分布式数据库等分布式存储系统[16]。

分布式存储系统最主要的原理就是分而治之，将系统分为多个自主的处理单元，在不同的节点上对数据进行存储和管理。分布式文件系统（DFS，Distributed File System）是基于客户机/服务机模式（C/S 模式）设计的，具有数据读取速度快、安全存储的特点。最常用的分布式文件系统有 GFS（Google File System）、HDFS（Hadoop Distributed File System）、TFS（Taobao File System）等[17]。

除了典型的分布式文件系统之外，还有各种非关系数据库分布式存储方案，常用的非关系式存储方案有 HBase、Redis 和 Mongodb 等。

2.5.3 数据处理

由于大数据区别于普通数据的特征还有特殊的业务需求，所以需要更为先进的计算分析方式。最为典型的计算方式有批处理计算模式和流处理计算模式[18]。

批处理（Batch Processing），即批量处理，先存储数据，后进行处理，主要应用于大规模的静态数据集，具有有界性、持久性以及大量性的特点。适用于对于时间要求不高的大量历史数据处理分析。

流处理（Streaming Processing）不同于批处理，是直接对数据进行处理。流处理中，将数据视为数据流，追求实时处理数据，并返回处理结果。所以，适用于对于时间要求较高的应用场景，如传感器网络、电力以及金融交易等。

目前主流的大数据处理框架有 Hadoop、Storm、Samza、Spark 和 Flink。这些框架中，Hadoop 仅支持批处理，Storm 和 Samze 支持流处理，而 Spark 和 Flink 则两种处理方式都支持。

2.5.4 数据分析

对比传统的数据分析，大数据对于数据分析无论是深度还是广度的要求都大大提升，因此传统的统计分析软件很难满足大数据的分析需求。为了克服传统的统计分析软件的扩展性较差，且 Hadoop 自带的分析功能较为薄弱的缺点，所以 IBM 公司提出了对 Hadoop 与 R 进行集成的方式，这样 Hadoop 就深度挖掘分析信息的能力[19]。此外，Python 也是目前主流的数据分析工具之一。

3 汽车行业大数据分析架构的需求分析

汽车行业由于其进行大数据应用所需数据的多样性以及数据源的特殊性，需要汽车行业的大数据分析平台有区分于其他行业的基础架构，以保证这些数据收集、存储以及分析能够顺利进行。

本文通过对各种类型企业大数据架构的研究，结合对目前大数据在汽车行业的应用分析，抽象出一种较为通用的大数据框架，来满足汽车行业对于海量数据的收集、存储以及分析需求。

本章的思路是，通过对大数据在汽车行业的应用归纳，总结出这些业务所需要的所有数据源，并分别分析这些数据源的特性及数据架构，进行归纳总结，提出汽车行业大数据分析平台基础架构的功能性需求。并根据软件架构设计的准则以及大数据的特性，提出非功能性需求。

3.1 数据源分析

未来，大数据在汽车行业的应用主要可以总结为九个方向[20]：

第一，市场营销：通过对消费者行为的分析，促进汽车的营销。具体实现为，通过社交网络，收集消费者的购买倾向、反馈等信息，来设计相应的营销主题来满足用户需求；

第二，智能预测：通过车联网实时传递的车辆数据，如可以远程分析出车辆的潜在风险，并及时警醒用户或者通过用户的实时坐标，分析出前方行驶路径的路况等；

第三，汽车保险：根据用户的驾驶行为和车辆信息，评估用户驾驶风险系数，并根据这个系数给用户提供相应的保险合同；

第四，评估二手车：目前，国内虽然很多消费者有二手车购买需求， 但是碍于市场上的二手车良莠不齐，虚假信息多，减少了消费者的购买欲望。但是在大数据的帮助下，通过车辆的历史维修数据等，可以合理地估计出二手车的价值；

第五，智能交通：大数据时代，也是汽车互联时代。未来，汽车通过互联网实时上传定位信息等，有助于缓解城市交通等

第六，汽车设计：通过收集用户舆情数据、分析司机使用报告，根据用户的偏好，设计符合用户需求的新功能；

第七，零部件采购：利用零部件采购数据，分析出原有供应链的缺陷，帮助制造商来优化采购流程；

第八，汽车制造：收集汽车制造过程中的数据，来进行制造模拟，优化流水作业线；

第九，财务优化：利用用户行为、偏好与购买力等信息，来发展更有效的财务项目，开展新的服务，从而有新的利润来源。

研究发现，大数据最常见的数据源有三类：机器、传统数据库以及社交网络。

机器产生的数据是机器数据。机器数据的英文是 Machine-generated Data，翻译成中文就是机器产生的数据。这类数据包括传感器数据、设备产生的日志数据、智能仪表产生的数据等。

传统数据库数据包括了绝大多数的企业数据库。企业常用的数据库有 ERP、SAP 等系统的数据库，都是传统数据库。传统的数据库就是关系式数据库，包括 Oracle、MySQL 等。

社交网络数据是社交网络产生的数据。举个例子微博、QQ 空间上的用户动态、评论等都算是社交网络数据。

根据这个分类标准，总结大数据在汽车行业的应用所需的数据以及数据源类型，见下表 3.1。

表 3.1 大数据在汽车行业的应用总结

Table 3.1 The summary of big data applications in automotive industry

应用	数据	数据源类型
市场营销	社交网络数据	社交网络
智能预测	车载传感器数据	机器
汽车保险	车载传感器数据、客户销售数据	机器、传统数据库
评估二手车	车辆维修数据、车辆定价数据	机器、传统数据库
智能交通	车载传感器数据	机器
汽车设计	车载传感器数据、社交网络数据	机器、社交网络
零部件采购	采购数据	传统数据库
汽车制造	制造数据	传统数据库
财务优化	财务相关数据	传统数据库

根据表 3.1,汽车行业的大数据应用覆盖了大数据最主要的三类数据源，包括社交网络、机器以及传统数据库。因此，汽车行业的大数据架构必须满足这三大类数据的需求。这三种数据源产生的数据特点不同，所需要的数据分析基础架构也不同。因此，具体分析这三类数据源的基础架构特性对于设计汽车行业的大数据架构十分重要。

3.2 大数据分析架构的功能性需求

大数据系统有两种架构观点：一是价值链架构观点，利用数据价值链，将大数据平台分为数据收集、数据存储、数据分析和应用。这种观点更适合用于产业界较为简单的大数据系统架构，所以本文的架构设计也是根据这个观点。二是层次架构观点，将整个大数据平台分为基础设施层、计算层和应用层。

根据价值链架构观点，大数据分析平台的基础功能包括数据收集、数据存储、数据分析与应用。因为汽车行业的大数据应用的主要数据覆盖了社交数据、机器数据以及传统数据库数据。所以，通过

以下具体分析这三类数据的收集、存储、分析需求，总结出其适用的数据分析架构，进而总结出车企大数据平台的功能性需求。

3.2.1 社交数据的数据分析架构

根据价值链架构观点，抽象出社交数据的数据分析架构如图 3.1 所示。

根据该数据分析架构，网络数据分析的基本流程为：

第一，网络数据抓取。因为网络数据多为半结构化或者非结构化数据，所以利用爬虫程序或者开源 API 从网页上抓取，并在本地存储；

第二，利用非关系式数据库存储抓取的数据。因为社交网站每天生成的数据量十分大，所以采用非关系式数据库存储，有利于读写效率；

第三，利用批处理配合统计分析工具进行社交数据的数据分析。因为社交网络数据最为常见的应用就是深度挖掘进行用户行为预测等应用。这种应用不需要时效性，但是对于数据挖掘的要求较高。所以采用批处理和统计分析工具结合的方式，能够较好地满足这种需求。

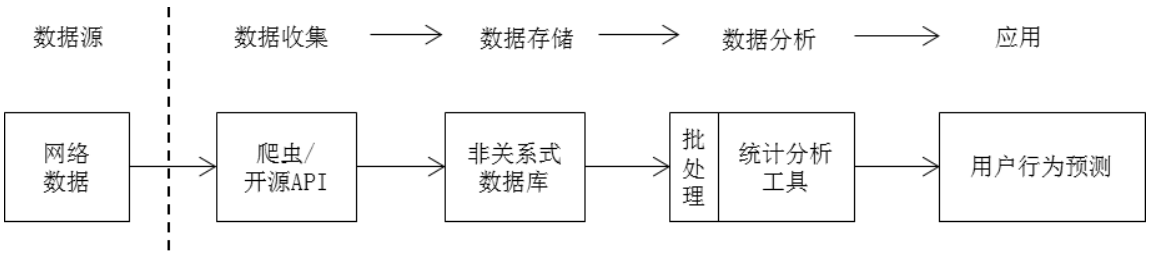

图 3.1 社交数据的数据分析架构

Fig.3.1 Data analytics architecture for social data

3.2.2 机器数据的数据分析架构

根据价值链架构观点，抽象出机器数据的数据分析架构如图 3.2 所示。

根据该数据分析架构，网络数据分析的基本流程为：

第一，传感器数据的收集。因为传感器数据的数据量极大，而且这些数据的时效性较强，所以采用日志采集工具能够保证以较快的速度不出错地采集数据；

第二，将数据存放在分布式文件系统中。因为车辆数据是实时生成的，所以传感器每时每刻都在接受着海量数据，分布式文件系统的容错能力极强，可以保证实时大规模数据传输不出错；

第三，利用流处理的方式配合统计分析工具进行机器数据的数据分析。车载传感器数据主要为车辆数据、用户行为数据以及环境数据等，这些最为常见的应用有实时路况提醒、汽车状态监控等。这

些应用对于时间的要求较强，要求系统能以较快的速度处理数据并返回结果。所以采用流处理和统计分析工具结合的方式，能够较好地满足这种需求。

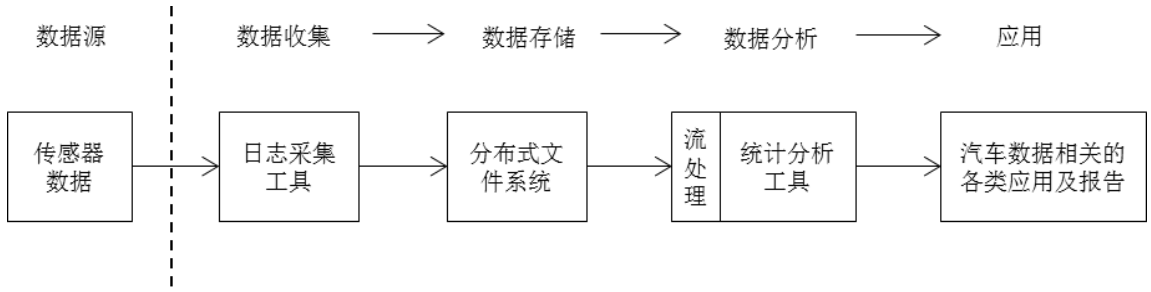

图 3.2 机器数据的数据分析架构

Fig.3.2 Data analytics architecture for machine-generated data

3.2.3 传统企业数据的数据分析架构

根据价值链架构观点，抽象出机器数据的数据分析架构如图 3.3 所示。

根据该数据分析架构，传统企业数据分析的基本流程为：

第一，传统数据的集成。企业有各种各样的信息系统，这些信息系统一般有各自的数据库。常见的企业信息系统通常都使用传统的关系式数据库 Oracle、MySQL 等。由于这些数据存储在不同的数据库中，数据格式多样，所以需要专门的异构数据集成工具，把业务需要的数据转存到数据分析平台的数据存储系统中；

第二，数据存储。传统企业数据库数据在经过数据集成后，具有格式统一、数据量较小的优势，所以对于数据的存储并没有太多具体要求。可视具体的业务需要而定，将这部分数据存储在分布式文件系统或非关系式数据库中，等待下一步分析；

第三，利用传统的统计分析工具进行分析。传统的企业数据由于数据量小，不需要用特定的大数据算法对数据进行处理再进行分析，所以传统的统计分析工具如：Excel、R 等就能满足这些分析需求。

由于大数据分析的需要，有很多大数据的应用可能不止需要一类数据源，如车辆保险预测，将会同时需要传感器收集的机器数据还有汽车销售数据这种传统的数据。这时，传统数据的分析将会和机器数据的分析结合在一起，所以就应该选择更为适应海量数据的数据存储方式配合大规模数据的计算框架来对这些数据进行分析。

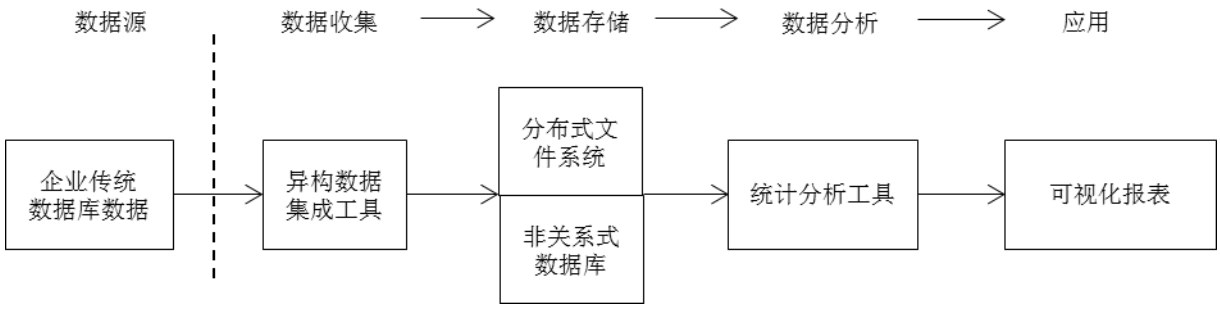

图 3.3 传统企业数据的数据分析架构

Fig.3.3 Data analytics architecture for traditional enterprise data

3.2.4 功能性需求总结

根据以上对于社交数据、机器数据以及传统企业数据的数据架构的分析，总结汽车企业大数据分析基础架构各个流程的功能性需求见下表 3.2。

表 3.2 汽车行业大数据基础架构的功能性需求

Table 3.2 Functional requirements of big data architecture for automotive industry

数据分析流程	功能性需求
数据收集	满足异构数据集成的需求；支持海量日志采集
数据存储	采用分布式存储的方式；支持非关系式数据库存储
数据分析	支持批处理；支持流处理；支持普通的统计分析
应用	能够满足多种应用需求；生成可视化保镖

3.3 大数据分析架构的非功能性需求

对于汽车行业大数据分析架构的非功能性需求分析以软件系统的非功能性需求为基础，并结合车企大数据的特，总结出车企大数据分析平台基础架构的非功能性需求有安全性、可扩展性和可靠性三点。

安全性：车企的大数据平台不仅存储了企业内部的采购、生产以及制造数据，还有用户的用车行为和车辆数据等。因此，系统的安全性尤为重要。

可扩展性：大数据分析平台基础架构的可扩展性非常重要。因为大数据时代代表着无限可能，大数据的应用不可能仅限于目前的几个方向，所以该基础架构应该能支持新的应用的接入。

可靠性：可靠性指的是在限定的条件和时间内完成规定功能的能力。对于大数据分析平台来说，进入车联网时代后，需要能够实时处理并分析大量的用户数据并给出实时报告，超出时效的报告意义不大。所以，系统的准确性尤为重要。

4 汽车行业大数据分析架构设计

本章承接上一章的需求分析，先根据功能性需求设计了，并根据功能性需求，选择了相应的大数据工具，设计了汽车行业大数据分析的逻辑架构方案。并具体分析了数据收集、数据存储、数据分析中每一组件的选择原因。最后，得出结论，该框架能够满足上一章中提出功能性需求以及非功能性需求。

4.1 汽车行业大数据分析的功能性架构设计

根据功能性需求的分析结果，设计汽车行业大数据分析的功能性架构如图 4.1 所示。该功能性架构是整合三类数据源数据分析架构的结果，能够满足这三类数据源收集、存储以及分析的需求。

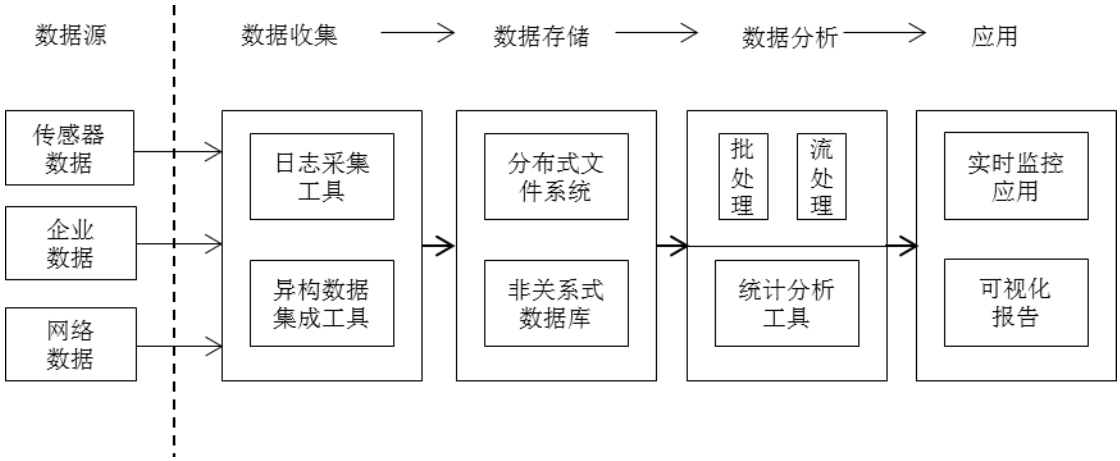

图 4.1 汽车行业大数据分析的功能性架构

Fig.4.1 Functional big data analytics architecture for automotive industry

4.2 汽车行业大数据分析的逻辑架构设计

根据汽车行业大数据分析的功能性架构，设计车企大数据分析平台的逻辑架构如图 4.2 所示。从各组件来看， 负责数据收集的有异构数据集成工具 Sqoop 和日志采集工具 Flume；数据存储则有分布式文件系统 HDFS 配合 Hive 数据仓库，并且支持非关系数据库 Cassandra；负责数据分析的有支持批处理的 MapReduce 算法，支持流处理的 Spark Streaming 和 Kafka，还有传统的统计分析工具 R 语言。这些工具的组合能够满足三大数据源的所有大数据应用的需求。

对于传感器数据来说，经过异构数据集成工具 Sqoop 集成，然后根据相应的需要进行存储分析。

对于网络数据来说，经过日志采集工具 Flume 收集后，存储到 HDFS 中，利用 Hive 就可以进行简单的查询，如果需要深层次的挖掘就可以利用 MapReduce 处理框架，或者配合使用 R 语言。

对于传感器数据来说，数据经由日志采集工具 Flume 收集后，存放到根据分析需要存放到 HDFS 中或者非关系式数据库 Cassandra 中。若是需要深度挖掘可以存放到 HDFS 中利用 MapReduce 结合 R 语言进行分析。若是需要实时分析，则可以利用 Cassandra、Kafka 与 Spark Streaming 组成的流处理框架进行分析。

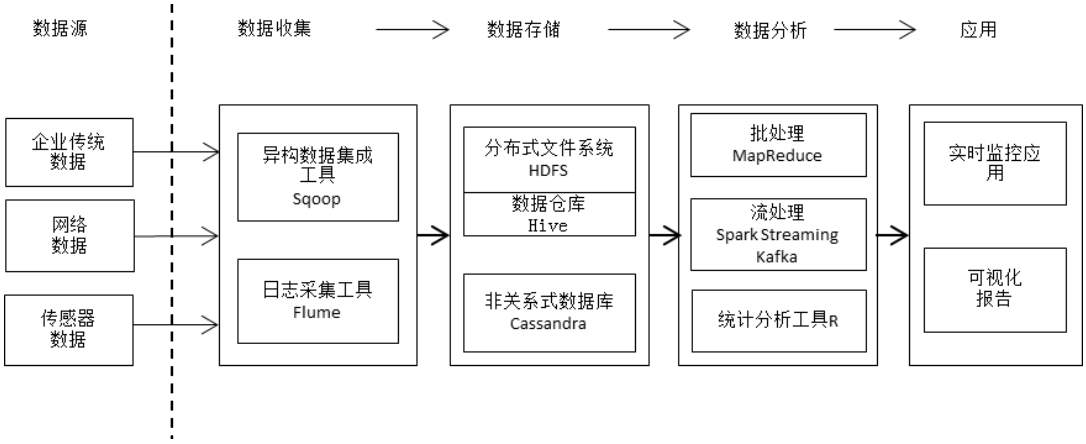

图 4.2 汽车行业大数据分析的逻辑架构

Fig.4.2 Big data analytics logical architecture for automotive industry

4.3 数据收集工具的选择

本架构选择 Sqoop 和 Flume 组合作为数据收集的工具，因为该组合能够满足几乎所有类型的数据的收集，包括结构化数据、非结构化数据与半结构化数据，并具有将异构数据整合并转存在本地数据库的功能。同时，Sqoop 和 Flume 的适应性都很强，二者的组合能与各类数据存储工具对接，包括各种分布式文件系统、数据仓库以及非关系式数据库。

Flume 是 Cloudera（该公司利用 Hadoop 这一开源软件基础架构开发自己的产品）基于 Hadoop 开发的用于数据收集与处理的系统。Flume 具有高可用性，高可靠性的特点。Flume 能够对数据进行简单处理，并写入到各种数据接受方。车企大数据数据分析平台基础架构中，Flume 用于收集网页服务器传来的各种日志信息并转存到 HDFS 中。

目前，最为流行的大数据环境下解决数据异构问题的迁移工具就是 Apache 旗下的 Sqoop。Sqoop 是一个独立的 Apache 项目，主要用于在 Hive 与传统的数据库之间的数据传递，可以将一个关系型数据库（如 ：MySQL ,Oracle 等）中的数据导入到 Hadoop 的 HDFS 中，或者将 HDFS 的数据导入到关系型数据库中。Sqoop 的优点有：专门为 Hadoop 设计，对于 Hadoop 的支持度极高；支持并行导入，宣称速度极快，支持按字段拆分并行化的导入过程；自带丰富的辅助工具；利用 MapReduce 计算框架，效率更高。

4.4 数据存储工具的选择

本架构采用分布式的存储方案，由分布式文件存储系统 HDFS 与分布式分关系数据库 Cassandra 相互配合。Hive 作为数据仓库，支持类 SQL 的 HQL 语言，能够从 HDFS 中抽取业务所需要的数据，进行查血。

Hadoop 是目前最为流行的大数据处理平台，从某种程度上说，Hadoop 已经成为大数据处理工具的标准[21]。用户在不了解分布式底层细节的情况下，可以利用 Hadoop 开发分布式程序，进行高效率的运算和存储。HDFS 是 Hadoop 的核心组件之一，是一种分布式文件系统，具有高容错性的特点，适用于价格低廉的硬件。同时，HDFS 能提供高吞吐量，访问应用程序的数据，十分适合大规模数据集的应用。HDFS 放宽了 POSIX（Portable Operating System Interface，可移植操作系统接口）的要求，可以实现以流的形式访问文件系统中的数据[22]。这些特点正适合车企所需要的处理大规模车辆数据的需求。

Hive 是一种建立在 Hadoop 上的数据仓库工具，具有学习成本低的特点，非常适合数据仓库的统计分析。它提供了一些可以用来进行 ETL（Extract, Transform, Load，提取转换加载）可以存储查询以及分析存储在 HDFS 中的大规模数据，并定义了简单的 SQL 查询功能（HQL）。

Cassandra 是 Facebook 开发的一个开源分布式的非关系（NoSql）数据库系统。它最初用于储存邮件收件箱等简单格式数据。2008 年，Facebook 将 Cassandra 开源，此后，由于其良好的可扩展性，被 Digg、Twitter 等知名网站所采纳，成为一种极为流行的分布式结构化数据存储方案。Cassandra 是非关系数据库中功能最多的，并且最像关系式数据库。

4.5 数据分析工具的选择

本架构所选择的数据分析工具 MapReduce、Spark Streaming、Kafka 和 R 能够满足不同类型，不同层次的分析需求。首先，最简单的就是传统的数据分析，只需要利用 R 就能够完成；其次，较为基础的大数据批量分析，利用 Hadoop 核心组件 MapReduce 就能完成；再者，深度的历史数据挖掘可以利用 R 与 Hadoop 集成，完成较为深层次的分析；最后，利用 Spark、Kafka 和 Cassandra 流处理框架能够满足事实分析的需求。以下，具体介绍流处理的原理。

Kafka 是一种分布式发布订阅消息系统，它也具有高吞吐量的特点，可以存储并且处理大规模网站中所有的动作流数据，这些动作包括网页浏览，搜索和其他用户的行动等。由于 Hadoop 即是一个日志数据和离线分析系统，又有实时处理的要求，Kafka 为其提供了一个良好的可行性解决方案。

Spark、Kafka 和 Cassandra 架构是当前最先进的快数据框架。所谓快数据框架，即能够同时满足流处理与批处理的需求，而 Hadoop 的核心组件 HDFS 只支持批处理。

这个架构处理数据的基础流程为，Spark Streaming 从 Kafka、数据库或者文件系统中获取数据后，利用 Spark 的全方位分析功能来对数据进行处理，最终存储到 Cassandra 中。对于一些无需实时处理

的历史数据，Spark 可以选择采用批处理，进行分析，而对与新收到的数据 Spark 则选择流处理的方式进行分析。在该框架中，Kafka 成为数据源与 Spark 之间的桥梁，获取流数据并输入到 Spark 进行处理分析。Cassandra 作为典型的非关系式数据库，特别适合存储 Spark 的分析结果。

4.6 总结

因为本章中提出的汽车行业大数据分析逻辑架构是根据功能性需求以及功能性架构提出的，显然，该架构能够满足汽车行业大数据分析架构的功能性需求。

汽车行业大数据分析架构的非功能性需求有安全性、可扩展性和可靠性。以下，具体分析该架构是否能够满足这三点需求。

安全性具体来说就是数据安全，所以需要考察所选择的数据存储方案的安全性。对于 Hadoop HDFS 来说，提供了保持数据安全的备份方式。而 Cassandra 则能够支持 TLS/SSL，能够保证用户的数据不泄露。所以，能够满足安全性需求。

可扩展性要求该基础架构能够快速适应新应用，该架构采用 Hadoop 与 Spark 整合的基础框架，能够适应多种应用的不同需求，所以具有较好的可扩展性。

可靠性要求数据处理的性能较高，并且在操作中不丢失数据。Hadoop 与 Spark 整合的框架，利用 Spark Streaming 算法，能以极高的效率处理数据，同时 Hadoop 有文件多备份的保障机制。因此，该架构能够满足可靠性的要求。

综上所述，该架构能够满足汽车行业大数据分析架构的功能性需求与非功能性需求。

5 奥迪中国大数据分析平台的基础架构设计

本章的主要目的在于验证上一章所提出的汽车行业大数据分析基础架构是否合理，在实践中是否能满足汽车企业的实际业务需求。

奥迪中国目前正处于企业转型的节点，迫切需要企业自己的大数据分析平台以开展相关的数据分析业务，为企业节约成本，提高利润。本文作者因为在奥迪中国信息技术部门实习，所以有机会加入到其大数据分析平台建设的项目中来，为其大数据分析平台的基础架构提供可行性方案。

本章的思路为分析奥迪中国目前的大数据分析基础架构，然后根据收集的奥迪中国的实际大数据架构需求，分析本文提出的架构理论是否能够满足这些需求，最后在奥迪中国应用该大数据分析架构。

5.1 目前的基础架构

对于实际企业大数据分析基础架构的设计，首先要满足企业原有的数据架构的功能，所以先考虑企业目前的数据分析相关业务运行状况。奥迪中国目前涉及到大数据分析的信息平台，只有一个：实时车辆监控平台（RTM Platform，Real-time Monitoring Platform），以车辆数据分析（Vehicle Data Analytics）为基础，该平台主要有以下两个功能：收集、存储并处理实时车辆数据流，并且能够实时在客户端显示结果；评估并处理存储下来的车辆数据，并且能以报表的形式在客户端显示。

其工作原理就是通过车载传感器实时收集的海量车辆数据通过日志采集工具采集后，存储在分布式文件系统中，根据所需的应用类型，选择流处理的分析方式，实现实时监控、错误分析、信号分析以及车辆分析等应用。其基础架构如图 5.1 所示。

<table>
<tr><td>应用</td><td colspan="2">汽车数据相关的各类应用及报告</td></tr>
<tr><td>数据分析</td><td>流处理
Storm</td><td>统计分析工具
Excel</td></tr>
<tr><td>数据存储</td><td colspan="2">分布式文件系统HDFS</td></tr>
<tr><td>数据收集</td><td colspan="2">日志采集工具Flume</td></tr>
<tr><td>数据源</td><td colspan="2">车辆数据</td></tr>
</table>

图 5.1 奥迪中国目前的大数据分析基础架构

Fig.5.1 Existing big data analytics architecture in Audi China

分析上图基础架构可知，奥迪中国目前的大数据分析基础架构实际上就是一个传感器数据分析的基础架构，并且不能支持深度挖掘，只是有较为基础的数据分析工具 Excel。

5.2 需求分析

根据对奥迪中国大数据分析平台的主要用户，包括企业管理者、业务部门、信息技术部门等进行访谈结合对其现有基础架构的分析，总结出奥迪中国大数据分析平台的需求见下表 5.1。

表 5.1 奥迪中国大数据分析平台的需求

Table 5.1 Requirements of big data architecture for Audi China

需求内容	需求来源	需求类型
1. 具有流分析的能力	现有基础架构	功能性需求
2. 能够集成企业现有其他信息平台的数据	业务部门	功能性需求
3. 数据能够安全存放	业务部门	非功能性需求
4. 能够支持深度分析	信息技术部门	功能性需求
5. 支持在该平台上增加新的应用	信息技术部门	非功能性需求
6. 具有原有 RTM 平台的功能	企业管理者	功能性需求

以下，具体分析奥迪中国大数据平台的各个需求，验证本文提出的大数据分析平台的基础架构是能够满足奥迪中国大数据分析架构的需求。

需求 1，具有流分析能力，本文提出的架构中的 Cassandra、Spark Streaming 以及 Kafka 组合能够满足这一需求。

需求 2，能够集成企业现有其他信息平台的数据，本文提出的架构中异构数据集成工具 Sqoop 能够满足这一需求。

需求 3，数据能够安全存放，这个本质上就是上文中提出的非功能性需求中的安全性需求，所以本文提出的架构能够满足这一需求。

需求 4，能够支持深度分析，本文提出的架构中的 Hadoop 与 R 集成，能够进行深度的数据挖掘，所以能够满足这一需求。

需求 5，支持在该平台上增加新的应用，这个本质上就是上文中提出的非功能性需求中的可扩展性需求，所以本文提出的架构能够满足这一需求。

需求 6，具有原有 RTM 平台的功能，上一节分析奥迪中国现有数据基础架构实际上就是传感器数据分析基础架构，而本文提出的架构设计已经满足的传感器数据集成的要求，所以能够满足这一需求。

综上所述，本文提出的架构能够满足奥迪中国大数据分析架构的所有需求，所以本架构可行。

5.3 奥迪中国数据分析平台基础架构设计

根据需求分析的结果，设计奥迪中国大数据分析平台的基础架构如图 5.2 所示。

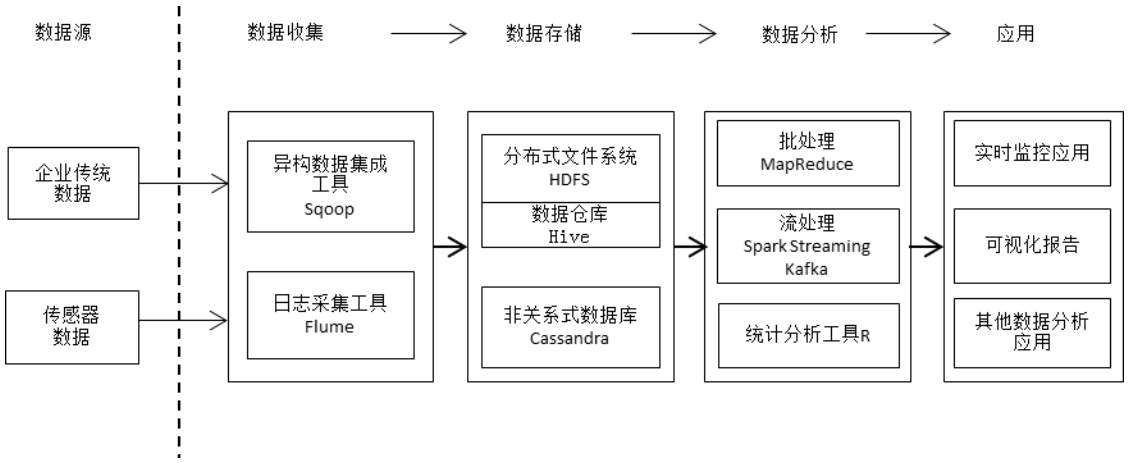

图 5.2 奥迪中国大数据分析架构

Fig.5.2 Big data analytics architecture for Audi China

该平台与车企的数据分析基础架构有一点不同：数据源没有网络数据。因为在奥迪中国提供的应用场景中，没有需要用到网络数据这一数据源的。因此就没有在架构中设计这部分。后期，如果对网络数据有需求的话，只需使用爬虫程序或者开源 API 抓取数据，本架构已经具备了分析网络数据所需要的所有功能。

5.4 结论

在奥迪中国成功应用本文提出的汽车行业的大数据分析架构，有以下三点意义：

第一，为奥迪中国成功设计了企大数据分析平台的基础架构，为其之后开展相关业务奠定了良好的基础；

第二，通过在奥迪中国应用汽车行业的大数据分析架构，证明了该架构不仅仅是一个理论空壳，而是一个实际可行的方案；

第三，通过在奥迪中国应用该架构，为日后其他汽车企业以及类似的企业建设自己的大数据平台提供了理论依据以及实际应用流程。

6 奥迪中国大数据实施管理方案

6.1 实施管理方案简介

目前，奥迪中国大数据分析平台的基础架构已经完善，但是没有任何具体的数据分析项目或应用在该平台上运行。因此，奥迪中国急需一个平台实施管理方案，尽快将平台投入使用，为企业创造受益。

根据实际需求，该实施管理方案的设计主要分为组织管理、业务流程两个部分。

组织管理部分，主要理清平台运营以及数据分析相关项目中所有角色的职责分配。因为由于奥迪中国没有该平台基础设施搭建所需要的具体环境，所以该平台现在搭建在大众集团的数据中心中，并交由大众集团的信息技术部门代为进行平台运营，所以理清各部门与角色的职责十分重要；

业务流程部分，主要设计一个新的数据分析项目或应用在该大数据分析平台运行的业务流程以及提供这一业务流程所需要的需求文件。据此，才能尽快将这个数据分析平台利用起来；

该方案的最终目的是希望该大数据分析平台能平稳运营，并为奥迪中国企业管理有限公司的管理层提供决策支持，达到降低企业成本，提高企业盈利的效果。

6.2 平台实施管理方案的设计框架

管理活动主要分为三个层级，战略层级、战术层级和操作层级。根据这一原则，并找出每个层级的关键元素，设计数据分析平台实施管理三层框架如图 6.1 所示。

战略层有三个关键要素，政府，企业，信息技术部门，这个由之前的需求分析可知。

战术层则是围绕着项目为中心，所以其核心就是项目管理，要考虑到项目管理的核心要素，成本、质量、时间、范围和项目干系人等。

操作层就是以数据分析平台为核心，该数据分析平台的基本流程就是数据收集、数据存储、数据处理、数据分析以及数据可视化。

而组织管理、业务流程和成本模型，则是贯穿始终的。所以，组织管理、业务流程设计离不开这三个层级。

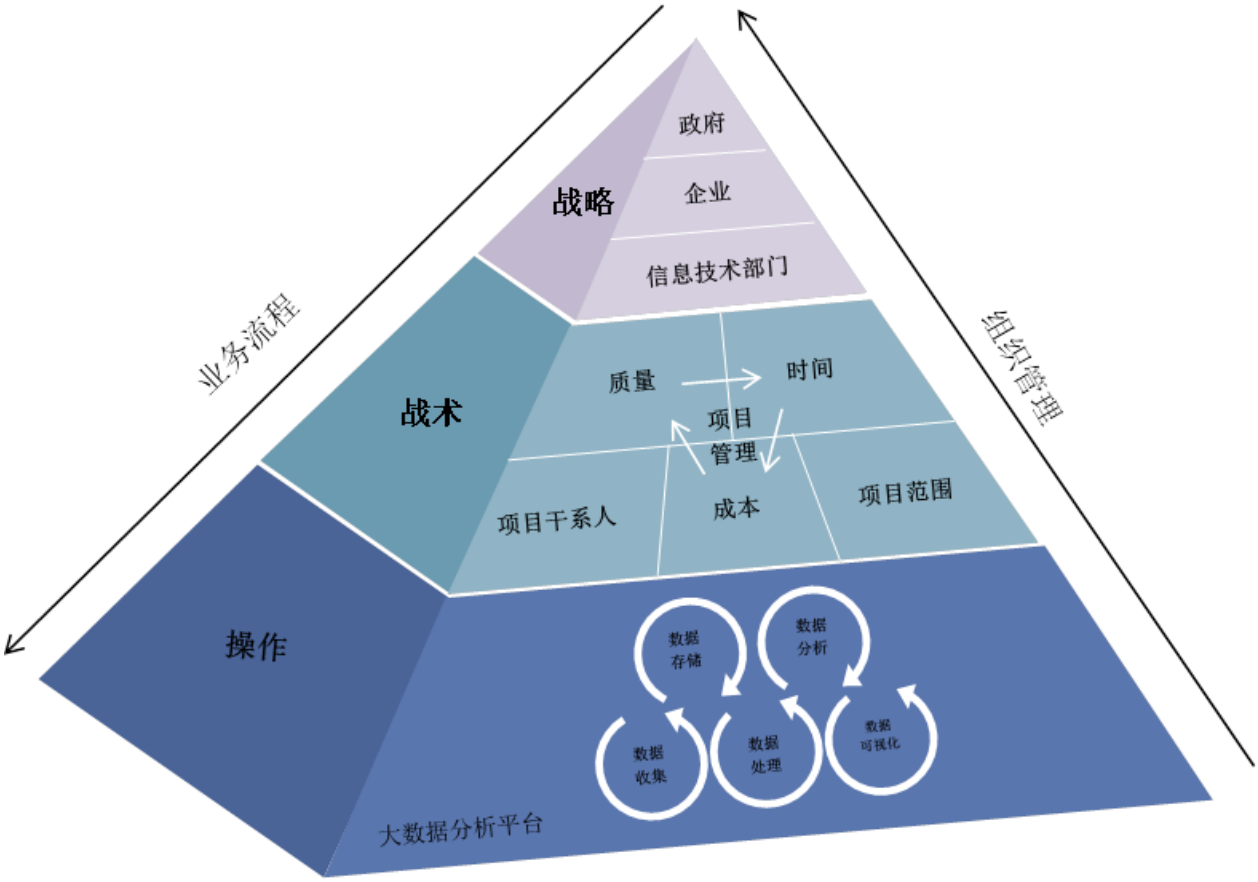

图 6.1 平台实施管理方案设计框架

Fig.6.1 The design framework of platform implementation and management solution

6.3 组织管理方案设计

6.3.1 组织及职责

奥迪中国大数据分析平台的正常运营应至少包含两大部分：第一，平台的日常运维，即维持目前已经与该数据分析平台连通的车辆实时数据的收集和分析。第二，在平台上增加新应用并正常运行。第一点，目前已经能够实现，奥迪中国大数据分析平台的基础设施现在交给大众集团的信息技术部门进行管理，并由他们外聘的服务供应商进行运维工作。第二点的实现则要涉及到具体的项目团队、采购流程以及财务流程。所以，平台的正常运营会涉及到的组织有：奥迪中国信息技术部门、大众集团信息技术部门、奥迪中国采购部门、奥迪中国财务部门和服务供应商。以下，具体分析各个组织的职责。

奥迪中国信息技术部门的主要职责就是监控和协调。监控，负责监控平台的日常运营和应用运转；协调，其他组织和部门并不直接接触，由奥迪中国信息技术部门负责所有的协调工作。

大众集团信息技术部门主要负责平台日常运营的监管。

奥迪中国采购部门主要负责增加新应用带来的采购流程。

奥迪中国财务部门这要负责增加新应用带来的财务流程。

服务供应商则负责项目的具体实施。这里会分为内部服务供应商和外部服务供应商。内部服务供应商是奥迪中国研发部门，外部服务供应商就是第三方服务供应商。

根据以上职责分配，以奥迪中国为核心，画出各组织关系如图 6.2 所示。

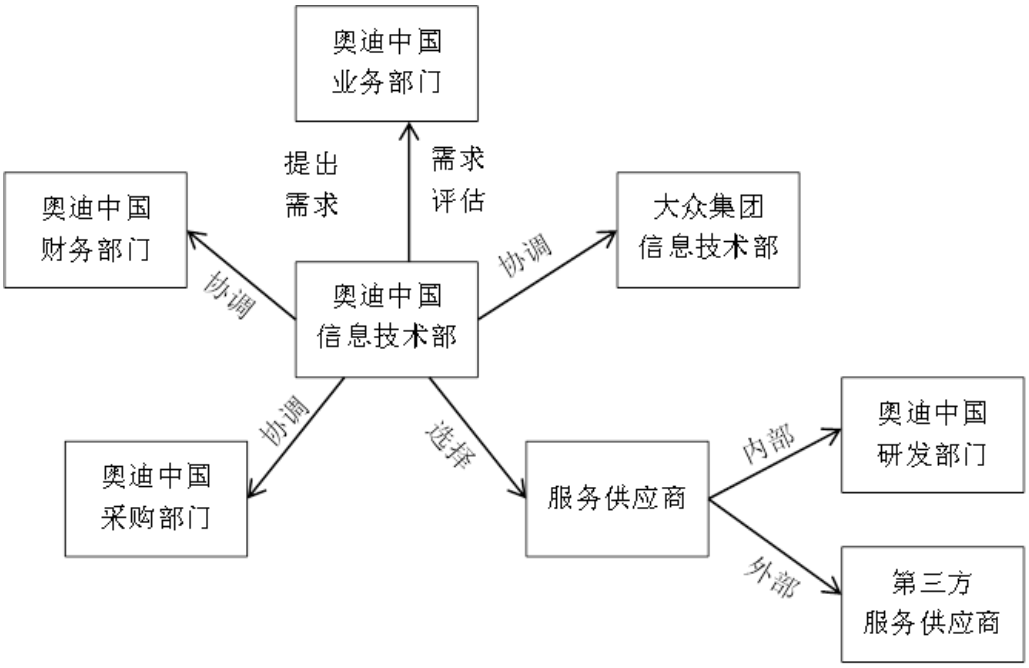

图 6.2 项目涉及的组织关系图

Fig.6.2 Platform-related organization chart

6.3.2 角色与功能

在理清了奥迪中国大数据分析平台管理活动的涉及到的所有组织之后，以下继续根据功能细分，围绕两个主要的运营活动，设立日常运维组与项目组。

根据现状，日常运维组主要是大众集团信息技术部门以及第三方服务供应商；

项目组则包括，奥迪中国信息技术部门、奥迪中国业务部门、奥迪中国采购部门、奥迪中国财务部门、大众集团信息技术部门和服务供应商。

所以，画出平台管理组织结构如图 6.3 所示：

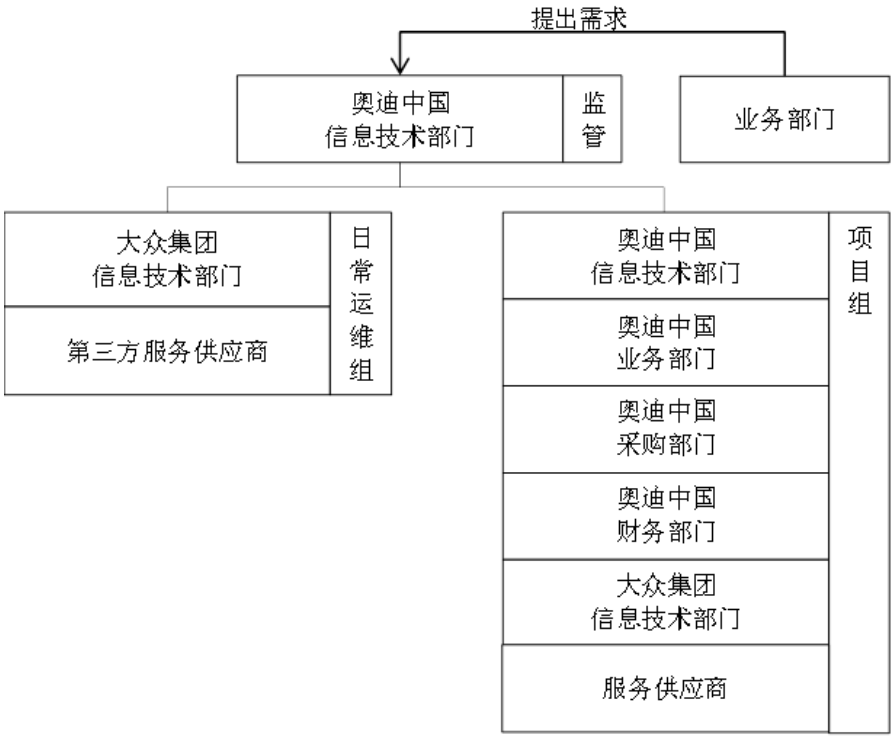

图 6.3 平台管理组织结构图

Fig.6.3 Organization chart of platform management

奥迪中国信息技术部门在这个管理体系中主要负责宏观监管两个团队的运作。

业务部门是项目的需求者，需要向奥迪中国信息技术部门提出需求申请，提交需求申请书，其中应该包括：项目背景与目标、市场大小和运营模型、项目参与方、数据需求和项目联系人。在提交需求申请后，业务部门有责任和项目经理一起优化项目方案。

以下，具体讨论日常运维组与项目组具体的成员及职责。

6.3.3 日常运维组

因为该实施管理方案主要是给奥迪中国信息技术部门提供的，而该平台目前主要交由大众集团信息技术部门管理。所以，以下并不对平台运营团队的职责分工进行详细描述。

平台运营团队主要有两个成员：大众集团信息技术部门和第三方服务供应商。

大众集团信息部门主要负责：宏观监管平台的运营状态还有选择第三方服务供应商。因为大众集团的信息部门的主要职能是项目管理，没有足够负责具体的技术支持的人员，所以 IT 基础设施运维的工作都是外包的。因此，需要选择第三方服务供应商。

第三方服务供应商：提供技术人员，如运维工程师等进行平台的运维。

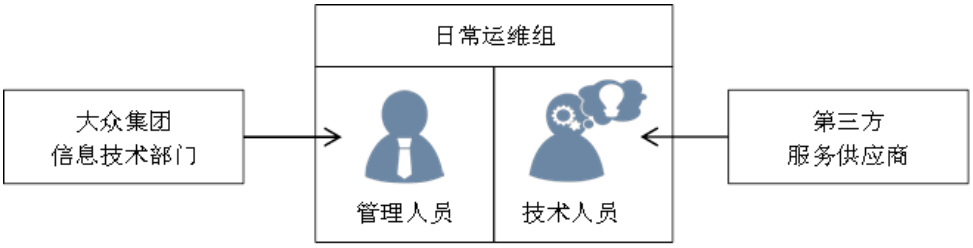

图 6.4 日常运维组

Fig.6.4 Daily operation team

6.3.4 项目组

项目组又根据具体的职责不同，分为项目实施团队和项目支持部门。

项目实施团队由奥迪中国信息技术部门、提出需求的业务部门还有服务供应商的成员共同组成。

奥迪中国信息技术部门：为每个项目指定项目经理。项目经理应负责项目管理的所有工作；负责其他项目干系人间协调，包括采购流程、财务流程等；为业务部门提供技术支持，包括提供 IT 预算、还有项目的技术需求等；选择服务供应商，根据不同的项目需求，可能为外部供应商或者内部供应商。

业务部门：指派员工加入项目实施团队，负责与业务需求相关的工作。

服务供应商：指派员工加入项目实施团队，根据业务部门的需求，负责项目的具体实施。

项目支持部门有大众集团信息技术部门、奥迪中国财务部门以及奥迪中国采购部门。

大众集团信息技术部门：在收到项目经理的新项目请求后，安排平台运维工程师给第三方供应商相应的项目成员授权，使其有权在平台上增加应用或者运行项目。

奥迪中国采购部门：奥迪中国企业管理有限公司的每个部门在采购部门都有一个对应的采购管理人员。由这个采购人员控制该项目的采购流程。因为这里面会涉及到第三方服务供应商的选择，需要购买服务，所以要生成采购订单等。

奥迪中国财务部门：奥迪中国企业管理有限公司的每个部门在财务部门都有一个对应的财务管理人员。由这个财务人员控制该项目的财务流程。因为这里面会涉及到对第三方服务供应商的付款问题。

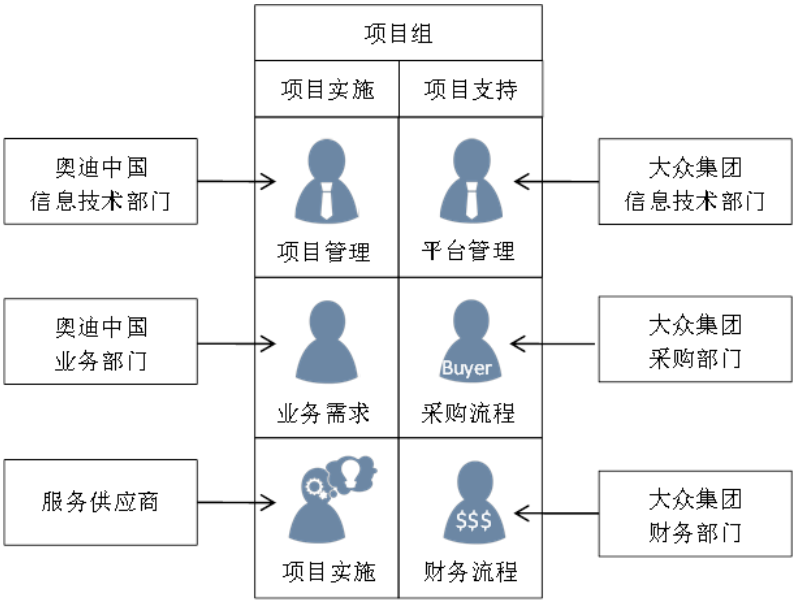

图 6.5 项目组

Fig.6.5 Project team

6.4 大数据应用项目实施流程设计

6.4.1 流程设计的必要性

目前，奥迪中国数据分析平台在大众集团的管理下进行日常运维，但没有任何项目及应用在该平台上运行。对于奥迪中国企业管理有限公司来说，该平台尚不能带来任何收入。奥迪中国信息技术部门已经收集了 18 个相关的项目需求，但是由于没有完善的实施管理方案以及业务流程，这些项目迟迟不能开展，这对于企业的运营是及其不利的。

尽快设计好项目在平台上上线的业务流程，可以有以下好处：

第一，提高项目管理的效率。在没有固定业务流程的情况下，项目管理容易走弯路，增加相关人员的工作量，对工作效率产生负面影响；

第二，吸引用户。在拥有完善的流程的基础上，用户在有了需求之后，可以便捷地找到相应的联系人，会更倾向于提出需求；

第三，增加企业收益。在企业中，所有活动的最终目的都是减少开支，增加受益。更多的大数据应用，可以便利企业各业务部门的工作，为管理层提供决策支持，最终为企业带来更多的受益。

6.4.2 流程设计目标

流程的设计目标得视需求而定。在数据分析的相关项目中，最主要的角色就是信息技术部门的项目经理或者项目团队还有项目的需求方。现阶段，综合各主要角色的需求，总结流程设计目标主要有如下两点：

第一，尽量简易的流程。因为奥迪中国信息技术部门的人手不足，需要一个项目经理同时负责多个项目，太繁琐的流程不利于项目的尽快上线；对于客户来说，繁琐的流程会降低他们提出需求的欲望；

第二，理清其中涉及的各部门的职责。因为目前该平台的管理交由大众集团，所以会涉及不同公司与部门之间的协调，分清各部门的职责，有利于项目的开展；

6.4.3 流程设计方案

因为奥迪中国管理企业管理有限公司都使用 GB1526-89，设计业务流程，所以沿用这种规范，设计数据分析项目上线的业务流程。

根据业务特点，将该流程分为五个阶段，下面对每个阶段进行简单的介绍：

阶段一，需求完善，该阶段涉及到的角色有需求发起人和奥迪中国信息技术部门，申请人向奥迪中国提交需求，并与信息技术部门一起完善项目需求书；

阶段二，采购过程，该阶段涉及到的角色有需求发起人、奥迪中国信息技术部门还有奥迪中国采购部门，该阶段需要三方达成项目采购流程的一致；

阶段三，项目准备，该阶段涉及到的角色有奥迪中国信息技术部门和大众集团信息技术部门，奥迪中国信息技术部门需要完成项目管理所需要的一切准备，而大众集团信息技术部门需要完成所有有关基础设施的技术准备；

阶段四，实施，该阶段涉及到的角色有奥迪中国信息技术部门和第三方服务供应商，该阶段第三方服务供应商完成项目的具体实施，奥迪中国信息技术部门检查项目质量；

阶段五，结束，该阶段这要涉及到的角色为奥迪中国信息技术部门，该阶段项目实施结束，转向项目的具体运作，

根据这五个阶段，画出具体的业务流程如图 6.6 所示。

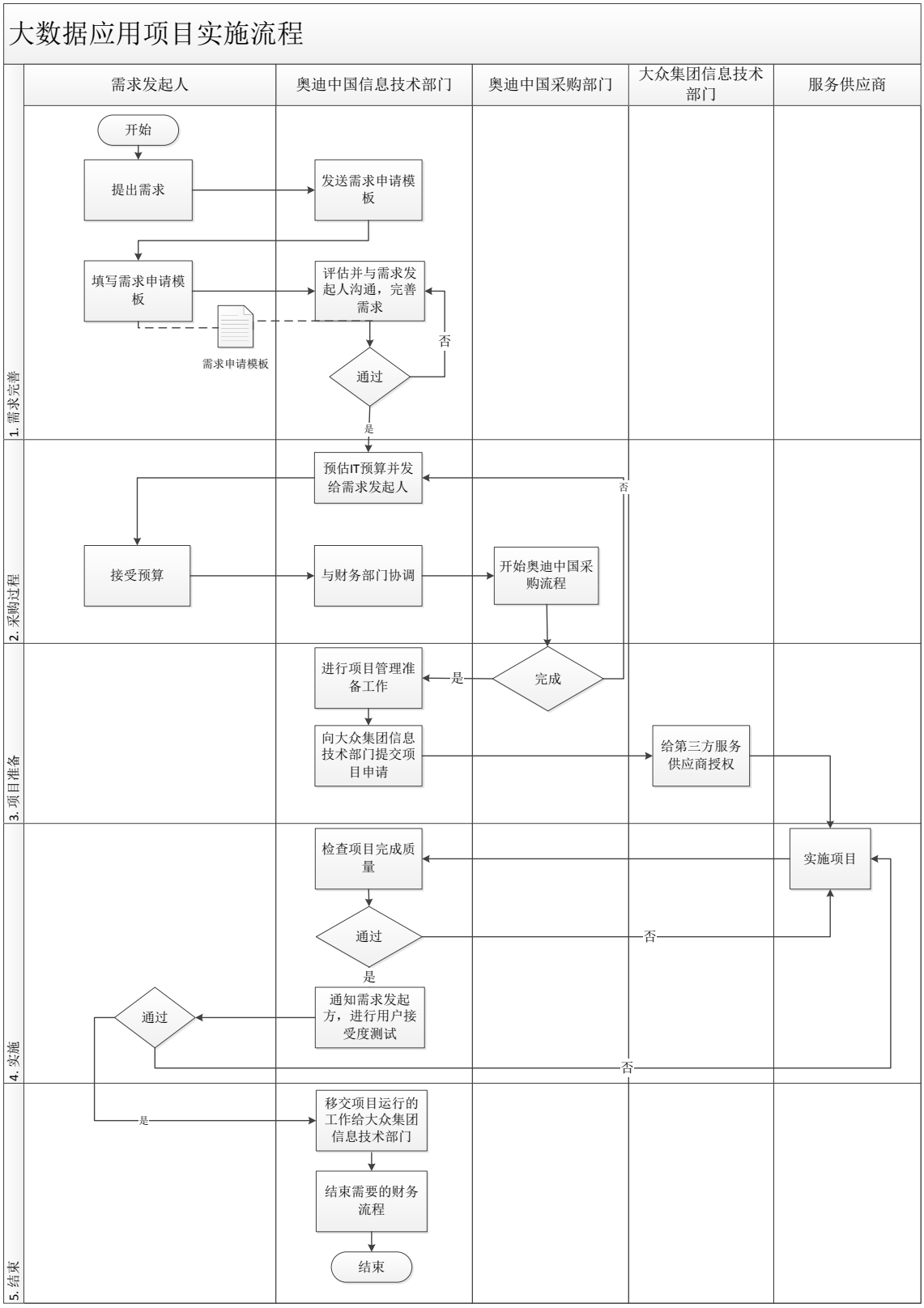

图 6.6 奥迪中国大数据应用项目实施流程图

Fig.6.6 Workflow of big data application implementation in Audi China

7 总结与展望

本论文从选题到最终定稿一共历时四个月。以下，从研究过程、研究成果以及展望三个部分，梳理整篇论文的脉络。研究过程按照论文的行文思路阐明了各阶段主要的研究内容；研究成果部分再次展示了本文最主要的研究成果，汽车行业的大数据分析架构；展望部分总结了本篇论文的不足之处以及未来的改进方向。

7.1 研究过程

本论文的研究过程分为以下几个步骤：

（1）提出问题。本文的第一章通过对于大数据背景及发展现状的分析，提出了设计汽车行业大数据架构的必要性；

（2）文献研究。本文的第二章通过分析国内外大数据研究现状、目前流行的大数据解决方案以及常用的大数据理论与技术，为本文研究的进行提供了理论基础；

（3）架构设计。本文的第三章和第四章是论文的主题部分，利用软件系统设计的方法进行需求分析，总结了汽车行业大数据分析架构的功能性需求和非功能性需求，并设计了汽车行业大数据分析的逻辑架构；

（4）架构应用。本文的第五章和第六章是对文中提出的汽车行业大数据分析架构的实际应用。通过在奥迪中国实际应用该理论模型，证实了该模型的可行性。

7.2 研究成果

本文最主要的研究成果就是汽车行业的大数据分析架构如下图 7.1 所示。

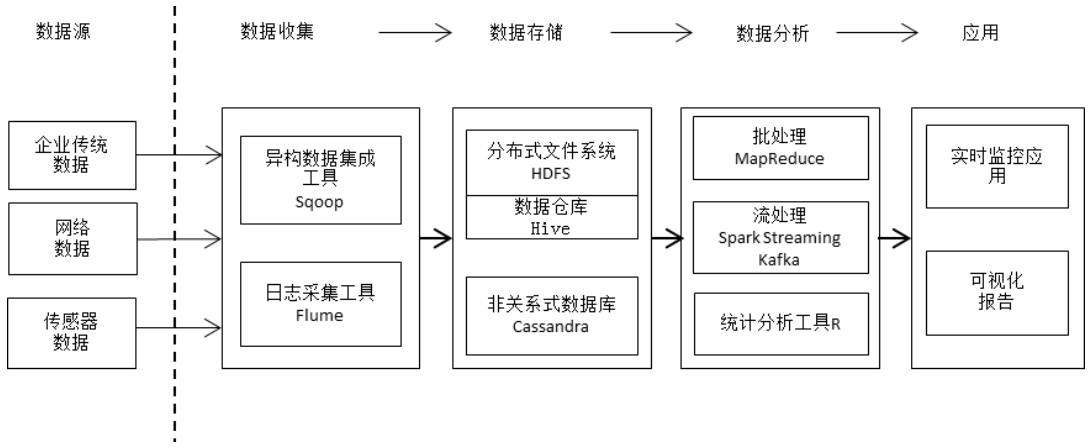

图 7.1 汽车行业的大数据分析架构

Fig.7.1 Big data analytics architecture for automotive industry

该架构的提出以及成功应用有以下四点意义：

第一，为今后传统车企进行转型，设计大数据分析架构、建设企业大数据分析平台提供了理论依据以及应用经验；

第二，汽车行业大数据分析架构的提出是一个理论创新，为今后其他学者继续汽车行业大数据分析架构研究提供了一定的基础以及可借鉴的方法；

第三，本文从业务角度入手，通过数据源分析的方式，进行大数据分析架构的设计，为今后其他学者设计其他行业的大数据分析架构提供了新的思考角度；

第四，该架构在奥迪中国的成功应用，推动了奥迪中国的大数据业务开展、数字化进程以及企业转型。

7.3 展望

作为一个本科生以及奥迪中国的实习生，由于本人的研究经验以及实际工作经验有限，本论文还有以下三点不足：

第一，对于汽车行业的大数据应用的九大方向是经由本人查阅资料总结出来的，所以可能还有遗漏，不能覆盖到所有的业务需求；

第二，由于本人的理论水平以及研究经验有限，只提出了汽车行业大数据分析的逻辑架构方案，没有涉及到各组件的具体实现；

第三，本文提出的实施管理方案只针对奥迪中国，对于其他企业没有太大的借鉴意义。

未来，希望能够继续学习，积累更多的实践经验，继续完善汽车行业的大数据分析架构以及实施方案。

致谢

本论文的顺利完稿，不仅仅是我个人的努力，更离不开多方的支持。

首先，最感谢我的导师李艳老师。从论文选题到最终定稿，老师不断解答我的各种疑问，并给我持续的支持与鼓励。没有李艳老师，我不能顺利完成这篇论文。

其次，我要感谢奥迪（中国），给我提供了在企业里完成毕业论文的机会，能够将理论与实践所结合。这里的同事，尤其是 Rolf Martinssen 和 Tuncay Iyilikli 为我对车企业务的深入了解提供了巨大帮助。

最后，感谢在论文撰写过程中所有帮助我支持我的老师、同学、同事、家人以及朋友。

感谢大家！

参考文献

[1] IDC.数字宇宙研究报告[R], 2012.

[2]佚名. 2017 年大数据分析预测[J]. 网络安全和信息化, 2017(4):6-6.http://old.ecas.cas.cn/xxkw/kbcd/201115_93655/ml/xxhjsyjcss/201212/t20121229_3730152.html.

[3]Ghemawat S, Gobioff H, Leung S-T. The Google file system [C]// Proceedings of the 19th ACM Symposium on Operating Systems Principles, 2003: 29-43.

[4]Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters [J]. Communications of the ACM, 2008, 51(1):107-113.

[5]Chang F, Dean J, Ghemawat S, et al. Bigtable: a distributed storage system for structured data[C]// Usenix Symposium on Operating Systems Design and Implementation. USENIX Association, 2006:15-15.

[6] Mcafee A, Brynjolfsson E. Big data: the management revolution[j]. Harvard Business Review, 2012, 90(10):60-68.

[7] HuLm Cui M, Branch H. Opportunities and Challenges of Big Data Development in Operators[J]. China Internet, 2014, 6(12):56-5.

[8] Diebold F X. On the Origin(s) and Development of the Term “Big Data”[J]. Pier Working Paper Archive, 2012 24(2):12-16.

[9] 李芬,朱志祥,刘盛辉. 大数据发展现状及面临的问题[J]. 西安邮电大学学报,2013,(05):100-103. [10] 汪云. 融合时代的大数据发展[J]. 电视技术, 2013, 3(22):1-3.

[11] 大数据应用浅谈[EB/OL]. https://wenku.baidu.com/view/94b19e08240c844768eaeeb5.html.

[12] IT 巨头厂商大数据解决方案[EB/OL].http://wenku.it168.com/redian/data/.

[13] 杨恒.软件架构设计[EB/OL].http://www.uml.org.cn/zjjs/201107193.asp.

[14] Divakar M, Shrikant K, Shweta J.Introduction of Big Data Classification and Architecture[EB/OL]. https://www.ibm.com/developerworks/cn/data/library/bd-archpatterns1/.

[15] 方巍,郑玉,徐江. 大数据:概念、技术及应用研究综述[J]. 南京信息工程大学学报(自然科学版),2014,(05):405-419.

[16]大数据存储技术进展[J]. 科研信息化技术与应用,2015,(01):18-28.现状及面临的问题[J]. 西安邮电大学学报, 2013, 12(5):100-103.

[17]中国计算机学会大数据专家委员会.中国大数据技术与产业发展白皮书[R].2013

[18]李贞强,陈康,武永卫,郑纬民. 大数据处理模式——系统结构,方法以及发展趋势[J]. 小型微型计算机系统,2015,(04):641-647.

[19]Das S, Sismanis Y, Beyer K S, et al. Ricardo: Integrating R and Hadoop [C]// Proceedings of the 2010 International Conference on Management of Data, 2010:987-998.

[20] 详解汽车行业大数据应用的九个大方向[EB/OL]. http://www.sohu.com/a/114259068_121534.

[21] 孟小峰, 慈祥.大数据管理: 概念、技术与挑战[J].计算机研究与发展, 2013, 50(1):146-169.

[22] HDFS Architecture Guide[EB/OL].[2014-08-25].http://hadoop.apache.org /docs/stable/hdfs-design.htm.