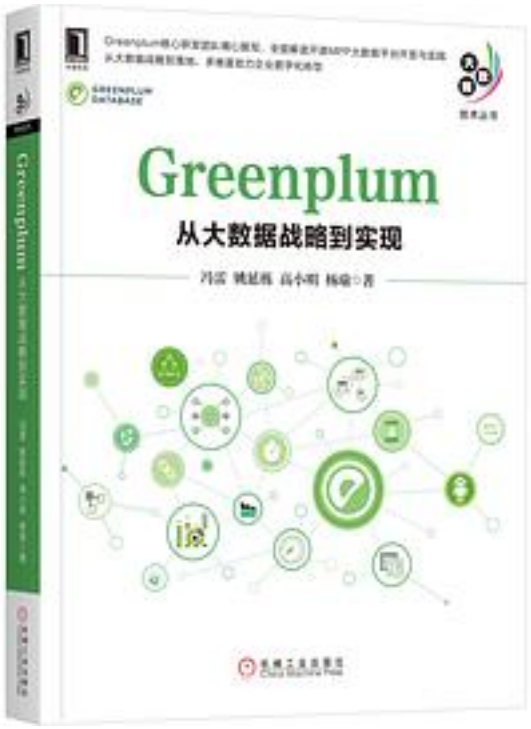
Greenplum：从大数据战略到实现



Greenplum：从大数据战略到实现_下载链接1_

著者:冯雷

出版者:机械工业出版社

出版时间:2019-7

装帧:平装

isbn:9787111632160

数字原生

2010年11月，在Greenplum创始人的支持下，我们在北京建立了Greenplum中国研发体系。2013年4月，随着Pivotal公司的建立，我们在Greenplum中国研发的基础上合并了部分VMWare中国研发集团的P层云资产，建立了Pivotal中国办公室。截至本书完稿的时候，我们的中国核心研发团队和全球研发团队一起奋斗了8年，打造的Cloud Foundry产品和Greenplum产品成为Pivotal公司在纽约证券交易所上市荣登PaaS第一股的基础。作为Pivotal中国办公室的创始团队，我们一直在审视和提升Pivotal中国办公室的使命和愿景。高尚的使命和愿景是促使一个机构达到世界一流水平的必要条件，因为

使命和愿景比战略更高一层。一个机构在前进的过程中，其战略不可避免地需要调整。在面对战略调整时，如果组织成员缺乏共同的使命和愿景，就很难在变化中存活下来。以PC行业为例，苹果公司由最初的苹果电脑公司（Apple Computers）发展到今天苹果（Apple）公司，业务也从以PC为重心迁移到以移动和云服务为重心。苹果公司的转型一路颠簸但最终成功，这与它们坚持艺术和科技的融合并提供一流的用户体验的使命是分不开的。对于不少没有完成转型的PC企业，仔细观察一下，会发现它们通常不能清楚地表达自己的使命。

那么Pivotal中国办公室的使命是什么？简单地说，是支持全球Pivotal产品和商业战略的成功。但是，这个回答显然不能说服和召集一批学霸把Pivotal中国办公室变成世界一流的创新机构。作者有幸参与Pivotal公司在EMC和VMWare内部的启动倡议（Pivotal Initiative），聆听到董事长Paul Maritz先生对Pivotal宣言（Manifesto）的解读。中国读者可能还不熟悉Maritz先生，根据维基百科的介绍，他是微软Windows平台的主要执行团队成员，负责过Windows 95和Windows NT等关键产品。在创建Pivotal之前，Maritz先生是VMWare公司的CEO，奠定了VMWare在虚拟化和I层云的行业领导地位。鉴于Maritz先生在业内的声望，作者仔仔细细阅读了他撰写的三页纸篇幅的Pivotal宣言，并且思考了Pivotal中国办公室如何既能拥抱Pivotal宣言又能在自己专注的领域成为国内意见领袖。今天，Pivotal的使命用一句话描述就是“The Way The Future Gets Built”，用中文直接翻译过来就是“构建未来的方式”。这句话显得有些抽象，所以在Pivotal中国办公室的日常事务中，我们会针对不同的团队来细化这句话：对于面向数字化转型客户的Pivotal Lab团队，这句话被表述为“交付一流的数字化转型体验”；对于云研发团队，这句话被表达为“通过Cloud Foundry云平台成为云原生平台的行业标杆”；对于数据库研发团队，这句话被阐述为“通过Greenplum成为大数据平台和机器学习的意见领袖”。这些使命背后的共同愿景就是提供“数字原生”世界的新产能，以及企业建立数字化所需要的软件平台和方法论。

数字原生就是从由物理世界为重心向数字世界为中心迁移时思考问题的方式。数字计算机发明之前，我们几乎没有什么数字资产和技术。数字计算机发明至今，我们对于数字资产的积累呈指数级增长，在我国更是呈现出跨越式发展的态势。举个例子，今天，如果我们出门不带手机，就会感觉寸步难行，本质上是因为手机已经成为我们进入数字世界的入口。通过手机，我们可以向数字世界发出各种请求，调度物理世界的资源为我们所用。Pivotal公司喜欢以“ask+综合部门@pivotal.io”的邮件方式来获得综合部门的支持。早期行政部门的同事刚加入Pivotal公司的时候常问我：“为什么不面对面请求，或者打个电话，又或者开个单子？”我的回答是这几种方式看似差别不大，但反映了思考问题方式的差别。Pivotal公司作为数字化的领导者，把软件和数据平台看作数字世界的入口。我们获取资源的方式是向这个数字世界发出请求。数字世界可能通过它的计算找到最优执行路径。有些工作的执行可能还需要转发给人进行人工处理，例如安装一台打印机。但是，有些请求则可以直接通过软件方式解决，例如申请一台云服务器。对于某些请求，虽然我们今天还无法完全以全数字化、无人干预的方式完成，但是，我们可以先把数字原生的框架奠定起来，为以后的进一步对接和持续改进做好准备。在作者看来，数字原生的持续改进过程分为三个阶段：

1）软件公司：通过数字应用实现数字世界和物理世界的无缝交互。

2）数据公司：通过大数据平台实现数据积累和数学模型运行支撑。

3）数学公司：通过数学模型的持续改进来最优化数字世界和物理世界资源。

因此，作者和团队希望能够以三部对应的著作（下面简称为“数字化三部曲”）在数字原生的征程上为读者提供战略参考和对应的软件平台及工具指导。

第一部曲：《Cloud Foundry：从数字化战略到实现》—这本书的主要目标是阐述企业如何实现数字原生第一阶段：实现数字化应用。该书讨论了云计算作为第三代技术平台带来的商业模式变更。在云计算的技术栈中，P层云带动了企业数字化浪潮。传统企业通过P层云可以迅速获得顶级互联网公司的软件迭代和发布速度，把与客户的交互通过消费级的应用数字化。书中例举福特公司通过FordPass建立了以汽车实体产品为核心的一系列用户数字化体验：汽车金融、远程监控车辆、停车位预留、旅途产品和服务推荐等。这个阶段也是一个持续改进的过程。以共享出行为例，今天用户通过手机平台进入数字世界，在打车应用中发送订单。打车平台通过选择最优执行路径，把订单发送给打车平台的司机。然后，司机在物理世界中驱车到达用户起点。随着有辅助的无人驾驶技术的成熟，这个数字世界的运行链条会继续延长，数字平台可以直接把无人车派送到用户起点。在其他的行业，数字应用的链条同样也在持续延长。

第二部曲：《Greenplum：从大数据战略到实现》（也就是本书）—我们的主要目的是阐述企业如何实现数字原生的第二阶段：大数据平台。随着数字应用的链条不断延长，企业需要一个大数据平台来积累应用生成的数据。这个工作听上去很容易，因为人们很早以前就使用磁带来存储数据，之后，存储媒介发生了巨大的变化，能够便捷地存储大量数据。那么为何还需要Greenplum这样一个大数据和机器学习平台？原因有两个：1）量大；2）快速计算。说到大，当数据量达到PB级别（相当于16000个64GB的iPhone中存储的数据）时，企业利用廉价但是可靠的存储来备份和管理是非常困难的。说到快，想象让用户从16000个iPhone的数据中寻找一张5年前的照片就可以感受到大海捞针般的困难；更何况企业的数据平台要支撑的机器学习和人工智能的数学模型的复杂度要比寻找一张照片的复杂度高几十到几万倍。可见，要想用极快的速度处理如此海量的数据是极其困难的。这也是企业在构建大数据平台时步履维艰的原因。Greenplum团队的优秀专家用企业积累了15年的知识和创新来解决这些难题：如何利用低价的存储设备来实现高可靠的数据存储？数据的存储如何为今天模型的计算做准备？如何给模型提供简单但又标准的接口？数据管理如何在“便于存储”和“便于日后查找”之间取得平衡？如何利用现在的I层云计算资源？如何访问文本和地理位置信息等各种数据源？如何访问和计算存储在其他系统（例如Hadoop）的数据？如何支撑今天主流的人工智能和机器学习模型？我们在创新过程中触碰到了很多计算机科学本身的极限。希望这本著作能给读者呈现一个解决了上述问题并可以实操的大数据平台和战略。

我们还在酝酿的第三部著作希望能帮助读者更好地实现数字原生的第三阶段：机器学习和人工智能。企业通过第一阶段和第二阶段的努力捕获和存储了大量的数据。为了更好地理解用户的需求，不少企业进入了更高阶的数字化战略：大数据驱动的机器学习和人工智能。在这个阶段的竞争中，企业会增设一个新的岗位：数据科学家。数据科学家会在大数据平台上创造和优化数学模型，以期待改进数字世界和物理世界的运作来更好地为人服务。前两部曲提供了软件工具和方法论以帮助企业成为基于大数据的人工智能和机器学习战略的数学公司，不少企业在实践过程中希望作者能够分享实践案例并就企业领导力转变提供咨询。考虑到这样一本著作的出版需要两年以上的时间，碰巧出版社和作者看到了顶级大数据咨询公司Booz Allen Hamilton的两位高管收集了大量实际案例的著作《The Mathematical Corporation: Where Machine Intelligence and Human Ingenuity Achieve the Impossible》，其中关于“数学公司”的提法和作者的观点不谋而合。通过出版社的努力，作者和团队把这部著作翻译成中文著作，可以作为第二部曲的伴侣著作来阅读。

虽然数字原生第三阶段的探讨还在创新者和早期用户者群体中进行，但是第二阶段大数据平台的建设已经在中国如火如荼地展开。大数据平台在数字原生三部曲中扮演了承上启下的关键角色，中大型的公司已经将大数据纳入信息平台的建设方案中。Greenplum因为开源生态和杰出的创新能力被列为方案的候选技术选项，这也使Pivotal中国办公室的同事们倍感欣慰。伴随Greenplum生态的持续发展壮大，希望这部著作能给企业高层制定战略提供建议和参考，既帮助工程团队开发应用，又能指导运营团队运维和保障。

本书内容组织方式

Greenplum经过15年的精心打磨，成为出色的开源MPP数据库和数据处理基础平台，已应用于银行、保险、证券、电信、物流、安保、零售、能源和广告等行业。我们希望本书能给已经建立或者准备建立大数据平台的企业决策者、架构师、开发人员、数据工程师、数据科学家和数据库管理员带来帮助，也希望从事大数据科研工作的教育工作者和学生能从中受益。

本书分为四个部分。

第一部分介绍大数据战略。其中，第1章将分享作者对于ABC（人工智能、大数据和云计算）之间关系的理解以及对人和人工智能的思考。第2章将介绍进取型企业为什么需要大数据战略以及如何建立大数据战略。

第二部分介绍大数据平台。其中，第3章将以数据平台演进历史和未来趋势为主题，描述三次整合的背景及影响，介绍选择大数据平台需要考虑的因素，以及为什么Greenplum是理想的大数据平台。第4章为Greenplum数据库快速入门指南。第5章将介绍Greenplum架构的主要特点和核心引擎。第6章将介绍数据加载、数据联邦和数据虚拟化。第7章将介绍Greenplum的资源管理以及对混合负载的支持。

第三部分介绍机器学习与数据分析。其中，第8章介绍Greenplum的各种过程化编程语言（用户自定义函数），用户可以使用Python、R、Java等语言实现用户自定义函数，还可以通过容器化技术实现自定义函数的安全性和隔离性。第9章将介绍Greenplum内建的机器学习库MADlib，数据科学家可以使用内建的50多种机器学习算法基于SQL对数据进行高级分析，并介绍如何扩展MADlib以实现新算法。第10章和第11章将分别介绍Greenplum如何对文本数据和时空数据（GIS）进行存储、计算和分析。第12章将介绍Greenplum丰富的图计算能力。

第四部分介绍运维管理和数据迁移。其中，第13章将介绍各种监控和管理工具及相关企业级产品。第14章介绍数据库备份、恢复和迁移。第15章和第16章将分别介绍如何从Oracle和Teradata迁移到Greenplum。

限于作者学识，本书难免有疏漏之处，恳请同行和各位读者批判指正，我们将不胜感激。您可以通过数字化三部曲的官网（DigitX.cn）或Greenplum中文官方社区（greenplum.cn）给我们留言并了解Greenplum的技术信息、获得著作的相关学习资源。

作者介绍:

冯雷(Ray Feng)

Pivotal中国常务董事(Managing Director)兼研发中心总经理。Pivotal中国成立至今，冯雷主持了近十亿人民币投资的中国运营和研发体系。作为Pivotal全球产品关键领导人，为Pivotal公司的数字化理念建立及其对应的Cloud Foundry和Greenplum产品提供战略输入。冯雷于2010年从美国硅谷归国，在世界500强公司EMC旗下组建了Pivotal中国。在归国之前，冯雷曾在500强企业甲骨文(Oracle)总部从事云计算产品研发。作为云计算早的一批从业人员，帮助甲骨文云计算资源调度领域成为意见领袖。拥有多项云计算专利。

姚延栋

Pivotal中国研发中心副总裁，在Pivotal公司全球范围内为Greenplum技术发展路线提供战略输入。联合创建了Pivotal中国研发中心，发起了Greenplum中国开源社区，奠定了包括阿里云、腾讯云和百度云在内的广大开源Greenplum用户群。在Pivotal中国招募并建设了Greenplum和HAWQ团队成为大数据和机器学习的意见领袖，培养团队成员同时

成为Apache和Greenplum代码提交者。在创建Greenplum/Pivotal中国之前, 曾在Sun Microsystem 与 Symantec
系统和存储部门工作多年。拥有多项国内外云计算和大数据专利。

高小明

Pivotal中国研发中心Greenplum产品总监，先后参与和负责数据分析协作平台Chorus、开源PaaS云平台Cloud
Foundry、MPP数据库Greenplum等产品的开发、运维和技术推广。目前着重关注PaaS云平台与大数据平台支撑下的数字化转型、微服务架构以及容器化与混合负载给数据产品带来的机遇和挑战。

杨瑜

Pivotal中国研发中心Greenplum工程技术总监，长期从事 Greenplum
内核的研发和管理工作，先后参与和负责基于Greenplum内核的机器学习库MADlib的研发、Greenplum
内核和PostgreSQL内核持续归并等工作，并参与组建Greenplum文本挖掘引擎GPText团队，有丰富的一线内核研发经验。

目录: 序

Greenplum：从大数据战略到实现_下载链接1_

标签

Greenplum

大数据

数字原生

数字化转型

好书，值得一读

postgresql

计算机

数据库

评论

战略部分深入浅出，观点论证严密，技术部分清晰、简练，跟着案例做一遍，收获大大滴

很好的一本书，看完之后对数据仓储，大数据和分布式等有了全新的理解和掌握

战略部分和后面的实现部分写的都很赞！

很多案例，步骤很清楚，通俗易懂

一直期待的书，既讲了战略又有实战，一口气读完，收获很大～

Greenplum入门看这本书就够了，进阶的话，还得仔细研究文档和源代码

一口气读完，对Greenplum有了更多的了解，很棒的一本书

作者写的很用心，对数据库和大数据平台有很深的思考

现在，很多企业都采用hadoop＋mpp架构，Greenplum是优秀的mpp平台，适合企业大数据

开源的技术越来越重要，作为目前唯一的大数据开源mpp平台，值得好好学习和了解，为未来做好准备

参加Greenplum的线下活动时知道要出版这本官方著作，都是Greenplum大神级的作者参与编写的，干货多多

强烈推荐，官方出品，内容权威，讲的很清楚

果然是核心研发人员出品，很多小细节都讲到了，对初学者很有帮助

现在，很多企业都采用hadoop＋mpp架构，Greenplum是优秀的mpp平台，适合企业大数据

刚拿到书，目录很全面详细，适合我，准备好好学习

第一章对大数据、人工智能、云计算的关系做了很全面的分析，特别是人工智能和人的关系的探讨，非常有深度，很受启发。

对如何应用大数据平台进行企业数字化转型有了全面的了解

的确是一本了解Greenplum的优秀作品

官方著作，对Greenplum这款优秀的mpp开源数据库进行了深度解读

作为Greenplum技术的爱好者，这本书让我对Greenplum有了更全面的认识。

书评

Greenplum：从大数据战略到实现_下载链接1_