

『赢在京东』系列技术教程

Presto

技术内幕

JD-Presto 研发团队 著

掌握大数据核心技术，
引领大数据实时计算潮流，京东技术大揭秘



中国工信出版集团

电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>



Presto 技术内幕

JD-Presto 研发团队 著

 JD.COM  京东大学  权威出品

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 介

Presto是专门为大数据实时查询计算而设计和开发的产品。由于Presto是基于Java语言开发的，因此，对使用者和开发者而言，Presto极易学习、使用并针对特定的业务场景进行改造开发和性能优化。无论是对多数据源支持，还是高性能、易用性、可扩展性等方面，Presto都是大数据实时查询计算产品中的佼佼者。

本书按照由浅入深的顺序对Presto进行了全方位的细致讲解，具体内容包括Presto概述、Presto安装与部署、Presto RESTful框架解析、提交查询、生成查询执行计划、查询调度、查询执行、队列、System Connector、Hive Connector、Kafka Connector、Connector开发、Functions开发、JD-Presto功能改造、Presto性能调优、Presto应用场景。

本书对Java研发工程师，在大数据技术领域从业的研发工程师、运维工程师、架构师，对大数据研究方向感兴趣的大学老师和学生，以及大数据技术的入门学者，从事海量数据分析应用开发者、研究者都有很强的参考价值。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

Presto 技术内幕 / JD-Presto 研发团队著. —北京：电子工业出版社，2016.7

“赢在京东”系列技术教程

ISBN 978-7-121-28563-9

I. ①P… II. ①J… III. ①数据处理—教材 IV. ①TP274

中国版本图书馆 CIP 数据核字(2016)第 073676 号

策划编辑：张慧敏

责任编辑：王 静

印 刷：三河市兴达印务有限公司

装 订：三河市兴达印务有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编：100036

开 本：787×1092 1/16 印张：19.5 字数：490 千字

版 次：2016 年 7 月第 1 版

印 次：2016 年 7 月第 1 次印刷

定 价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888，88258888。

质量投诉请发邮件至 zltz@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819 faq@phei.com.cn。

序言 1

12 年来，京东一直以开源技术作为构建自身核心技术体系的基础，在开源领域深耕多年，积极推动行业分享与交流。JD-Presto 是京东首个贡献于开源社区的软件产品，开创了京东回馈开源社区的先河。

一年前，JD-Presto 研发团队是京东众多研发团队中的普通一员，时至今日，他们已成长为京东负有盛名的研发团队之一。一年来，他们忍受着孤独与寂寞，面对着无数困难与挫折，但他们从未放弃，凭着对技术的热爱、对信念的坚守、对优秀产品的执着追求，勇敢面对困难与挑战，历尽艰辛，终于迎来胜利的曙光。目前 JD-Presto 在京东内部已成功应用于精准营销、安全识别、商家后台、财务报表等十多个生产系统。京东为拥有 JD-Presto 研发团队而骄傲！

与书店里的 Hadoop 书籍不同的是，本书是市面上第一本详细介绍 Presto 的书籍。它由浅入深地详细介绍了 Presto 的安装过程、内部运行原理机制、功能特性、性能优化方法，以及在应用过程中常见的问题及解决方案等，同时详细介绍了 JD-Presto 版本的许多新功能及其适用的业务场景。这些功能的成功研发，使得 JD-Presto 版本不仅可用于离线大数据分析计算来提升性能，还可用于多种数据源混合进行实时大数据分析计算，使一些使用其他大数据技术不能解决的业务场景有了一个全新有效的解决方案。同时，他们维护着 Presto 在中国社区的建设，鼓舞着和培养了一群 Presto 的爱好者。

今天，国内已有越来越多的技术人员与公司在使用的 JD-Presto 版本，相信本书能对读者学习和使用 JD-Presto 版本提供很大的帮助。本书不仅仅满足了初学者对技术书籍的渴求，也适合对 JD-Presto 版本有一定使用经验的人员学习参考使用。希望它能成为你在 Presto 技术旅程中的一个朋友，点亮你前进的道路。

张晨
京东集团 CTO

序言 2

记得我刚加入京东的时候，Hadoop 的集群建设方兴未艾，大数据计算纷纷被转到 Hadoop 中进行处理。然而，以前做数据挖掘的同学并不十分精通 Java 语言。他们更擅长于业务和 SQL 语言。为了让他们的工作更有效率、更得心应手，有必要找寻一个像 Google Dremel 一样，以 SQL 为驱动的交互式 Ad-hoc 查询工具。于是我花了一些时间进行评估，包括 Hive、Spark、Impala 等，后来 Facebook 刚开源不久的 Presto 便映入眼帘。Presto 设计简单精巧，可以处理海量数据，最大化地利用硬件性能，计算全部在内存中完成，很好地利用高速网络来进行数据调度，其编码风格一致，易于掌握和改进。虽然在当时 Presto 的数据源和 SQL 的支持度还有待丰富，但已经初步展示了发展的潜力，很适合京东的应用。Presto 就这样在京东找到了家。JD-Presto 研发团队具有多年 JAVA 编程和分布式并行处理经验，他们工作兢兢业业，在原先开源的基础上，丰富了多种数据源的处理接口，并且通过对任务调度的追踪和分析，改进并提升了并行处理任务的吞吐量。我们积极和 Facebook 的 Presto 团队合作，成为 Presto 在国内的首批贡献者，建立并运行 Presto 的中国社区。我们也和 CSDN 的同仁建立了 Presto 的开发者论坛，在国内积极推广，普及 Presto 知识，让更多的开发人员受益。

京东是开源社区的受益者，也是开源开放文化的拥护者。我们将自己取得的点点滴滴毫无保留地回馈于开源社区，与大家共享，和大家一起进步。为了更好地服务于对 Presto 有兴趣的研发人员，京东的同事们将平时工作中的积累和认识编辑成书，奉献给大家，希望他们的努力有助于读者学习和掌握 Presto。在此，我特别感谢为此书出版而辛勤工作的同事，他们是吕信、郭李明、袁安峰、孔云龙、戴东东等人。

翁志

京东集团首席技术顾问

序言 3

Presto is an open source, highly scalable, distributed SQL engine we developed at Facebook. Since releasing it two years ago, we have witnessed a tremendous amount of community interest and adoption. In the last twelve months, the Presto community has grown from technology firms in Silicon Valley to a broad variety of organizations worldwide. Today, NASDAQ, the second largest stock exchange in the United States, runs Presto to power their analytical use cases.

（Presto 是我们在 Facebook 开发的开源、高度可扩展的分布式 SQL 引擎。Presto 在两年前发布后，大量的社区开始关注和后续应用 Presto。在最近的 12 个月里，Presto 社区已经从一家硅谷科技公司成长为全球性的生态圈，涵盖一批多样化的组织。如今，全美第二大连交所纳斯达克也在使用 Presto 来驱动他们的案例分析业务。）

As organizations adopt Presto, many are contributing back to it. Teradata, a leader in data warehousing technology, maintains a group of engineers who contribute to Presto. We have also collaborated with engineers at JD.com since 2014 to make Presto better. Presto earned an "Open Source Outstanding Project" award at the 2015 China Open Source World Summit.

（组织机构们不仅在应用 Presto，他们中的一部分也在回馈 Presto。Teradata，数据仓库技术领域内的引领者之一，拥有一支专门为 Presto 贡献力量的工程师团队。从 2014 年起，我们与中国电商领域引领者京东公司的开发人员展开合作，精益求精地打造 Presto 产品。随后 Presto 在 2015 年“第十届开源中国开源世界高峰论坛”上获取“2015 COPU 开源优秀项目大奖”。）

We hope you will learn more about Presto through this book developed by engineers at JD.com, and we encourage you to work with the community to improve Presto.

（我们真诚地希望你通过这本由京东工程师们撰写的书更深入地了解 Presto 引擎，也鼓励并期待你携手社区，改善 Presto 引擎。）

Jay Tang

Manager of Presto product, Big Data, Facebook

前言

为什么要写这本书

相信很多软件开发人员在遇到应用系统涉及需要多种数据源、多种字符集下的混合计算时，都被数据的准确性和查询性能所困扰，一直以来都没有一个有效的技术方案能解决这个问题。2014年7月，京东CTO体系首席技术顾问翁志先生把Presto引入京东，数据准确性验证表明，Presto在多种数据源、多种字符集下的混合计算均表现优异，数据准确性符合财务标准。我们认识到，Presto是一个非常优秀的产品，尽管它有许多不足，我们决心把它打造为一个介于T+1计算和生产实时报表之间混合计算的大数据产品，支持秒级、分钟级获取查询结果、支持多并发的交互式Ad-Hoc查询与分析。Presto是一个较为完美的大数据实时计算解决方案，而市面上这类产品太少了，我们将经过生产环境验证的JD-Presto版本开源给业内使用，主导Presto中国社区的推广工作。应广大Presto技术爱好者的要求，我们编写了这本书，以帮助更多的初学者、大数据领域的从业者学习及使用JD-Presto。

和其他许多调研过Presto的部门或公司一样，我们最初搭建的Presto测试集群在京东这种大数据量的环境下复杂计算任务几乎运行不起来，JOIN查询也比较弱，集群时常宕机。同时，在测试过程中我们也发现，与Spark、Impala相比，Presto使用更为简单、易用、高效，这让我们深感振奋，一致认为这个产品值得团队花大力气去研究和优化。因此，我们专门申请了上百台高性能服务器来搭建性能测试集群环境，随后在几百TB级数据量下进行大规模的测试与验证，完全证实了Presto的平均性能是Hive等离线数据分析产品性能的10倍，并且支持ANSI标准SQL、数据源完全解耦等一系列优势。在之后一年多的时间中，我们研发了众多实用功能，着力修复了许多Bug，大幅提升了Presto的查询性能和并发执行性能，如重写了数据库驱动，使得Presto连接数据库进行海量数据查询的性能提升了几百倍；优化了内存使用效率，使查询性能对内存大小的依赖大幅降低；支持分布式数据库的分库分表设计；优化数据存储算法。通过上述一系列新功能的成功研发和对性能的大幅改进，使得JD-Presto的性能稳定是Hive的10倍以上，是Spark的3倍以上，集群成本节省一半，在一个大规模的大数据集群中，这种技术优势可带来显著的经济效益。

时至今日，京东已有近二十多个系统在使用 JD-Presto 版本，尤其是在精准营销平台中 JD-Presto 作为大数据即席运算平台，起到了关键性作用，极大提升了采销部门进行精准营销活动的效果和效率。在团队的努力下，有越来越多的公司使用了我们的产品，取得良好的社会效益并扩大了京东的技术影响力。

JD-Presto 研发团队先后荣获多个奖项：在“第十届开源中国开源世界高峰论坛”上荣获“2015 COPU 开源优秀项目奖”；在“第十四届中日韩三国 IT 局长 OSS 会议及东北亚开源软件推进论坛”上荣获“优秀项目技术奖”；在京东第四届 CTO 体系“金项奖”中荣获第一名。JD-Presto 产品蕴含着团队中这群不羁的兄弟们的心血、激情和梦想，从做产品研发决定的那一刻起，就注定我们不平凡。

本书适合的读者对象

- Java 研发工程师；
- 在大数据技术领域从业的研发工程师、运维工程师、架构师；
- 对大数据研究方向感兴趣的大学老师和学生，以及大数据技术的入门学者；
- 从事海量数据分析应用的开发者、研究者。

如何阅读本书

本书按照由浅入深的顺序对大数据实时计算产品 Presto 进行了全方位的细致讲解，从基本的安装部署与使用到源码解析，以及根据实际业务需求进行新功能改造开发和性能优化。

若你从来没有使用过 Presto，那么完全可以按照由浅入深的顺序从头开始阅读此书。

若你已经成功地使用了 Presto 一段时间，想要深入地了解 Presto 的设计思想并希望从源码级别了解 Presto 的实现思路，那么可以直接从核心设计篇进行阅读。

若你本来就是 Presto 的贡献者，苦于 Presto 不能满足你的实际业务问题，想要对 Presto 进行功能改造但是还没有明确的思路；又或者你在使用 Presto 的过程中发现其性能没有达到你的预期，而又找不到优化的方案，那么高级篇会给你提供一些建议。

在本书最后的附录部分不仅对 Presto 使用过程中出现的各种问题给出了明确的解决方案，还对 Presto 中的各个配置参数的含义和作用进行了详细的说明并给出了推荐配置值。

相信此书会为广大大数据爱好者提供一个新的视角，思考并解决大数据相关问题。

勘误和支持

由于作者水平和能力有限，编写时间仓促，本书存在不妥之处在所难免，我们真诚希望同行和广大读者朋友们不吝赐教。另外，如果你有关于 JD-Presto 的任何问题，可以加入我们的 QQ 群：141254058，与我们在线交流，或者访问网站：<http://prestodb-china.com/>，期待你的反馈意见，我们将不胜感激。

致谢

感谢现任京东 CTO 体系首席技术顾问、信息安全部、京东硅谷研发中心负责人翁志先生，在产品研发和推广使用的过程中给予了我们极大的支持和鼓励。同时在本书出版的过程中，从选题、审稿到出版无不得到他的热心帮助，在此致以深深的谢意！

感谢京东云平台——数据研发运营部的负责人樊建刚先生给予我们充分的授权和支持，他鼓励我们将产品开源、组建技术交流群和中文社区，使得 JD-Presto 成为京东第一个开源的软件产品，开创京东回馈开源社区的先河。

感谢研发团队的兄弟们，他们酷爱技术、注重团队协作、深入理解客户需求，能从一个较为全面的角度考虑产品的架构设计与功能研发，这是产品成功的基石，也是最关键的因素。

回首这一路的艰辛，无以表达此时的心情，只能说，兄弟们所做的一切都让我心怀崇敬！

戴东东

2016 年 4 月

写在前面

IT 领域中的某些概念可能真的很难用中文表述清楚其英文的含义，因此为了保证读者能够准确理解某些关键术语和概念的含义，在本书中对于一些关键的概念或者名词就不再翻译为对应的中文了。

由于本书中某些章节对源码进行了解析说明，本书中的所有源码都是针对于 Presto-0.107 进行讲解的，并且在对源码或者原理讲解时，仅写出了关键源码，对于非关键源码，没有进行书写。

目 录

第一部分 基础篇

第 1 章 Presto 概述.....	2
1.1 Presto 背景及发展.....	2
1.2 Presto 特点.....	2
1.3 基本概念.....	3
1.3.1 Presto 服务进程	3
1.3.2 Presto 模型	4
1.3.3 Presto 查询执行模型	5
1.4 Presto 整体架构.....	9
1.4.1 硬件架构	9
1.4.2 软件架构	9
1.5 小结.....	11
第 2 章 Presto 安装与部署.....	12
2.1 环境说明.....	12
2.2 准备工作.....	13
2.2.1 建立 SSH 信任关系	13
2.2.2 安装 Java	17
2.2.3 安装 Maven	17
2.2.4 安装 Hive.....	18
2.3 源码编译.....	21
2.3.1 下载源码	21
2.3.2 源码结构说明	22

2.3.3 编译	26
2.4 部署	30
2.4.1 服务部署	30
2.4.2 客户端部署	38
2.4.3 JDBC 使用	40
2.5 小结	42

第二部分 核心设计篇

第 3 章 Presto RESTful 框架解析	44
3.1 Statement 服务接口	44
3.2 Query 服务接口	47
3.3 Stage 服务接口	48
3.4 Task 服务接口	49
3.5 小结	52
第 4 章 提交查询	53
4.1 提交查询的步骤	53
4.2 源码解析	53
4.3 提交查询的流程	60
4.4 小结	61
第 5 章 生成查询执行计划	62
5.1 基本概念	63
5.1.1 Node	63
5.1.2 Metadata API	67
5.2 词法与语法分析	68
5.2.1 语法规则	69
5.2.2 词法分析	69
5.2.3 语法分析	71
5.3 获取 QueryExecution	72
5.3.1 获取 QueryExecutionFactory	72
5.3.2 创建 QueryExecution	73
5.3.3 启动 QueryExecution	74
5.4 语义分析	77
5.4.1 Statement 分析	77
5.4.2 Relation 分析	84
5.4.3 表达式分析	91

5.5	执行计划生成.....	91
5.5.1	执行计划节点	91
5.5.2	SQL 执行计划	93
5.5.3	Relation 执行计划	95
5.5.4	Query 执行计划	99
5.6	执行计划优化.....	102
5.6.1	ImplementSampleAsFilter	102
5.6.2	CanonicalizeExpressions	102
5.6.3	SimplifyExpressions	102
5.6.4	UnaliasSymbolReferences	103
5.6.5	PruneRedundantProjections.....	103
5.6.6	SetFlatteningOptimizer	103
5.6.7	LimitPushDown	104
5.6.8	PredicatePushDown	104
5.6.9	MergeProjections	104
5.6.10	ProjectionPushDown	104
5.6.11	IndexJoinOptimizer.....	105
5.6.12	CountConstantOptimizer	105
5.6.13	WindowFilterPushDown	105
5.6.14	HashGenerationOptimizer	105
5.6.15	PruneUnreferencedOutputs.....	106
5.6.16	MetadataQueryOptimizer	106
5.6.17	SingleDistinctOptimizer	106
5.6.18	BeginTableWrite	106
5.6.19	AddExchanges	107
5.6.20	PickLayout	107
5.7	执行计划分段.....	107
5.7.1	Source	107
5.7.2	Fixed	107
5.7.3	Single	107
5.7.4	Coordinator_only	107
5.8	示例.....	108
5.8.1	Count 执行计划.....	108
5.8.2	Join 执行计划.....	108
5.9	小结.....	110

第 6 章 查询调度	111
6.1 生成调度执行器	111
6.2 查询调度过程	113
6.2.1 NodeManager	114
6.2.2 NodeSelector	115
6.3 小结	118
第 7 章 查询执行	119
7.1 查询执行逻辑	120
7.2 Task 调度	120
7.2.1 Source Task 调度	120
7.2.2 Fixed Task 调度	126
7.2.3 Single Task 调度	128
7.2.4 Coordinator_Only Task 调度	128
7.3 Task 执行	129
7.3.1 创建 Task	129
7.3.2 更新 Task	135
7.3.3 运行 Task	140
7.4 小结	147
第 8 章 队列	148
8.1 配置说明	148
8.1.1 queues 队列定义	149
8.1.2 rules 规则定义	149
8.2 队列加载	150
8.3 队列匹配	151
8.4 小结	154
第 9 章 System Connector	155
9.1 System Connector 使用	155
9.1.1 Information_schema	155
9.1.2 Metadata	157
9.1.3 Runtime	157
9.2 System Connector 实现	159
9.2.1 Information_schema 实现	160
9.2.2 System Connector 实现	163
9.3 小结	168

第 10 章	Hive Connector	169
10.1	与 Hive 的结合	170
10.2	Split 分片管理	175
10.3	数据读取	179
10.4	Create Table As Select 的实现	182
10.5	小结	186
第 11 章	Kafka Connector	187
11.1	认识 Kafka Connector	187
11.1.1	配置	187
11.1.2	配置属性	187
11.1.3	内置字段	189
11.1.4	表定义文件	190
11.1.5	Kafka 中的 key 和 message	191
11.1.6	行解码	192
11.1.7	日期和时间解码器	194
11.1.8	文本解码器	194
11.1.9	数值解码器	194
11.2	Kafka 连接器使用教程	194
11.2.1	安装 Apache Kafka	195
11.2.2	下载数据	195
11.2.3	在 Presto 中配置 Kafka topics	197
11.2.4	基本数据查询	197
11.2.5	添加表定义文件	199
11.2.6	将 message 中所有值映射到不同列	200
11.2.7	使用实时数据	202
11.3	Kafka Connector 获取数据	207
11.3.1	Split 分片管理	207
11.3.2	数据读取	209
11.4	小结	210
第 12 章	Connector 开发	211
12.1	创建 Maven 工程	211
12.2	注册 Plugin	213
12.3	Connector	213
12.4	Metadata	215
12.5	SplitManager	217

12.6	RecordSetProvider.....	218
12.7	小结	219
第 13 章	Functions 开发.....	220
13.1	Function 注册	220
13.2	窗口函数	225
13.3	聚合函数	229
13.4	小结	232

第三部分 高级篇

第 14 章	JD-Presto 功能改造.....	234
14.1	PDBO 功能开发.....	234
14.1.1	JDBC Split 剖析	235
14.1.2	JdbcRecordCursor 剖析	238
14.1.3	分批次读取实现原理	240
14.1.4	动态步长实现原理	243
14.1.5	条件下发	245
14.1.6	PDBO 配置定义	247
14.2	DDL 及 DML 支持	250
14.2.1	Hive 连接器 Insert 功能	250
14.2.2	Hive 连接器 CTAS 动态分区表功能	252
14.3	动态增加、修改、删除 Catalog.....	254
14.3.1	目的	254
14.3.2	现状	254
14.3.3	实现	255
14.3.4	效果	258
14.4	小结	258
第 15 章	Presto 性能调优.....	259
15.1	合理设计分区	259
15.2	Group By 字句优化	259
15.3	使用模糊聚合函数	259
15.4	合并多条 Like 子句为一条 regexp_like 子句.....	260
15.5	大表放在 Join 子句左边.....	260
15.6	关闭 distributed hash join.....	261
15.7	使用 ORC 存储	261
15.8	小结	262

第 16 章 Presto 应用场景	263
16.1 ETL	263
16.2 实时数据计算.....	264
16.3 Ad-Hoc 查询.....	266
16.4 实时数据流分析.....	266
16.5 小结.....	268
附录 A 常见问题及解决办法	269
A.1 同时访问两个 Hadoop 集群.....	269
A.2 Kafka 集群重启后无法获取数据	272
A.3 Task exceeded max memory size.....	277
A.4 SQL 中 In 子句太长导致栈溢出错误.....	278
A.5 高并发导致大量查询出错.....	279
附录 B Presto 配置参数说明	282
附录 C Presto 执行信息说明	289