

# 实例讲解spark在京东智能供应链预测系统的应用

转载

2018年04月25日 18:08:37

166

问题导读：

1. 京东的供应链是什么样的呢？
2. 预测技术在京东的供应链起着什么样的作用呢？
3. 京东整个供应链的架构是什么样的呢？
4. 预测系统不同层面的技术选型分别为什么？
5. 预测系统以机器学习算法为主的核心流程是什么呢？
6. 预测系统以时间序列为主的核心流程是什么呢？
7. spark在预测核心层的应用是什么呢？

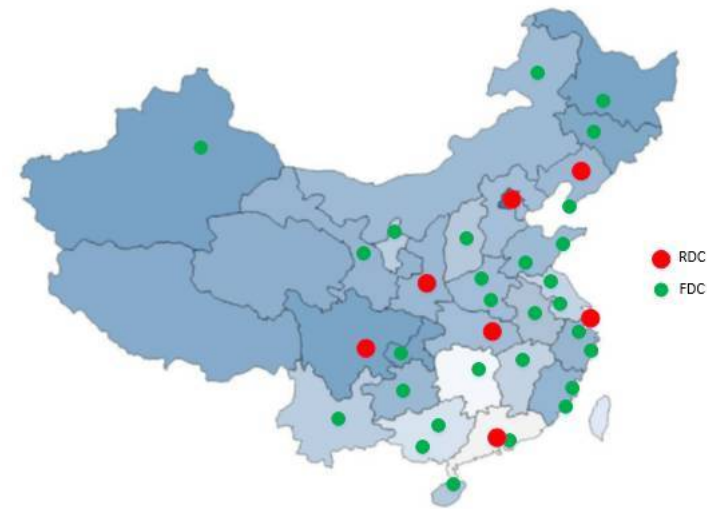
## 1. 背景

前段时间京东公开了面向第二个十二年的战略规划，表示京东将全面走向技术化，大力发展人工智能和机器人自动化技术，将过去传统方式构筑的优势全面升级。京东Y事业部顺势成立，该事业部将以服务泛零售为核心，着重智能供应能力的打造，核心使命是利用人工智能技术来驱动零售革新。

### 1.1 京东的供应链

京东一直致力于通过互联网电商建立需求侧与供给侧的精准、高效匹配，供应链管理是零售联调中的核心能力，是零售平台能力的关键体现，也是供应商与京东紧密合作的纽带，更是未来京东智能化商业体布局中的核心环节。

目前京东在全国范围内的运营256个大型仓库，按功能可划分为RDC、FDC、大件中心仓、大件卫星仓、图书仓和城市仓等等。RDC（Regional Distribution Center）即区域分发中心，可理解为一级仓库，向供货商采购的商品会优先送往这里，一般设置在中心城市，覆盖范围大。FDC（Forward Distribution Center）即区域运转中心，可理解为二级仓库，覆盖一些中、小型城市及边远地区，通常会根据需求将商品从RDC调配过来。



结合人工智能、大数据等技术，京东首先从供货商那里合理采购定量的商品到RDC，再根据实际需求调配到FD



原创

24

粉丝

13

喜欢

6

等级：博客 4

访问量：10

积分：1453

排名：3万



## 大数据可视化



### 博主最新文章

- 日志记录原则和方法
- Minimax算法及实例分析
- ML中相似性度量和距离的计算&n实现
- VS工程中出现白色文件，怎么移
- MFC使用CEF并实现js与C++交互解决Render进程中OnContextCre与OnWebKitInitialized的js扩展问题

### 文章分类

- C++
- VC
- htmlui
- python
- 嵌入式开发
- Linux

展开

### 文章存档

- 2018年3月
- 2018年2月
- 2018年1月
- 2017年12月
- 2017年11月

### 1.2 京东供应链优化

用户体验提升的同时也伴随着大量资金的投入和成本的提高，成本必须得到控制，整个体系才能发挥出最大的价值，于是对供应链的优化就显得至关重要了。

京东自打建立仓储体系的那一天起，就不断地进行改进和优化，并且努力深入到供应链的每一个环节。优化其实是一门运筹学问题，需考虑在各种决策目标之间如何平衡以达到最大收益，在这个过程中需要考虑很多问题，把这些考虑清楚，问题就容易解决了。举几个简单的例子：

- 商品补货：在什么时间，给哪个RDC采购什么商品，采购量是多少？
- 商品调拨：考虑在什么时间，给哪个FDC调配什么商品，调配量是多少？
- 仓储运营：在大促来临之际，仓库和配送站要增配多少人手、多少辆货车？

虽然看上去这些问题都很容易回答，但仔细想想却又很难给出答案，原因就在于想要做到精确不是那么容易的事情，就拿补货来说，补的太多会增加库存成本，补的太少会增加缺货成本，只有合理的补货量才能做到成本最低。

### 1.3 预测技术在京东供应链的作用

借助机器学习、大数据等相关技术，京东在很多供应链优化问题上都已经实现系统化，由系统自动给出优化建议，并与生产系统相连接，实现全流程自动化。在这里有一项技术起着至关重要的低层支撑作用--预测技术。据粗略估算，1%的预测准确度的提升可以节约数倍的运营成本。

怎样理解预测在供应链优化中的作用呢?拿商品补货举例，一家公司为了保证库房不缺货，可能会频繁的从供货商那里补充大量商品，这样做虽然不会缺货，但可能会造成更多卖不出去的商品积压在仓库中，从而使商品的周转率降低，库存成本增加。反之，这家公司有可能为了追求零库存而补很少的商品，但这就可能出现严重的缺货问题，从而使现货率降低，严重影响用户体验，缺货成本增加。于是问题就来了，要补多少商品才合适，什么时间补货，这就需要权衡考虑了，最终目的是要使库存成本和缺货成本达到一个平衡。

考虑一下极端情况，等库存降到零时再去补货，这时供货商接到补货通知后将货物运往仓库。但是这么做有个问题，因为运送过程需要时间，这段时间库房就缺货了。那怎么办呢?就是利用预测技术。利用预测我们可以计算出未来商品在途的这段时间里销量大概是多少，然后我们让仓库保证这个量，低于这个量就给供货商下达补货通知，于是问题得以解决。总而言之，预测技术在这里发挥了重要的作用，成为关键的一个环。

## 2. 京东预测系统

### 2.1 预测系统介绍

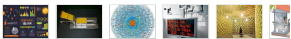


#### 博主热门文章

- 以色列Aladdin HASP SRM(AES 密狗破解经验分享)  
5511
- 日志系统框架的设计与实现  
4988
- APP中https证书有效性验证引发 (例Fiddler可抓https包)  
3367
- Window内存详解（五）VMM/看工具  
3296
- 如何用C#编写文本编辑器  
3213
- C# 开源库大全  
2933
- 国内地图坐标系介绍及常见地图高德、凯立德之间的坐标系转  
2723
- C++使用缓存加速文件的读取  
2305
- Sublime Text3 3143 以前注册问题  
2233
- 容器中查找最大值所在的位置  
2197



#### 新型燃料



#### 联系我们



- 请扫描二维码联系
- webmaster@csdn.net
- 400-660-0111
- QQ客服

关于 招聘 广告服务

©1999-2018 CSDN版权所有  
京ICP证09002463号

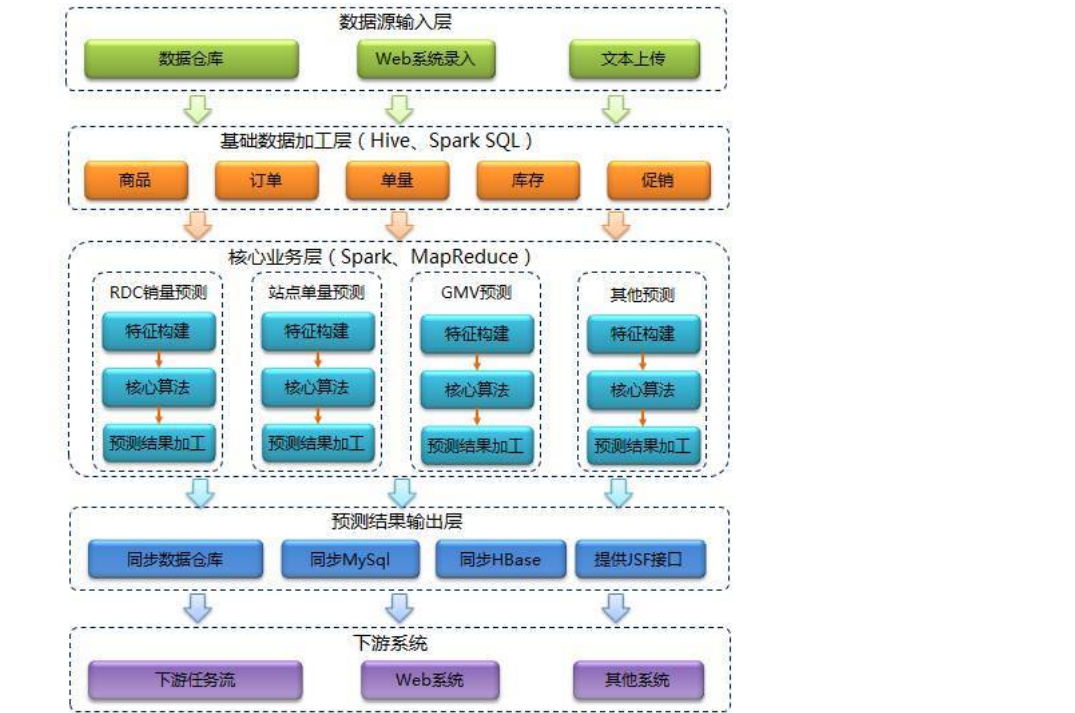
经营性网站备案信息  
网络110报警服务  
中国互联网举报中心  
北京互联网违法和不良信息举报中心

预测系统在整个供应链体系中处在最底层并且起到一个支撑的作用，支持上层的多个决策优化系统，而这些决策优化系统利用精准的预测数据结合运筹学技术得出最优的决策，并将结果提供给更上层的业务执行系统或是业务方直接使用。

目前，预测系统支持三大业务：销量预测、单量预测和GMV预测。其中销量预测主要支持商品补货、商品调拨；单量预测主要支持仓库、站点的运营管理；GMV预测主要支持销售部门计划的定制。

销量预测按照维度又可以分为RDC采购预测、FDC调拨预测、城市仓调拨预测、大建仓补货预测、全球购销量预测和图销预测等；单量预测又可分为库房单量预测、配送中心单量预测和配送站单量预测等（在这里“单量”并非指用户所下订单的量，而是将订单拆单后流转到仓库中的单量。例如一个用户的订单中包括3件物品，其中两个大件品和一个小件品，在京东的供应链环节中可能会将其中两个大件品组成一个单投放到大件仓中，而将那个小件单独一个单投放到小件仓中，单量指的是拆单后的量）；GMV预测支持到商品粒度。

2.2 预测系统架构



整体架构从上至下依次是：数据源输入层、基础数据加工层、核心业务层、数据输出层和下游系统。首先从外部数据源获取我们所需的业务数据，然后对基础数据进行加工清洗，再通过时间序列、机器学习等人工智能技术对数据进行处理分析，最后计算出预测结果并通过多种途径推送给下游系统使用。

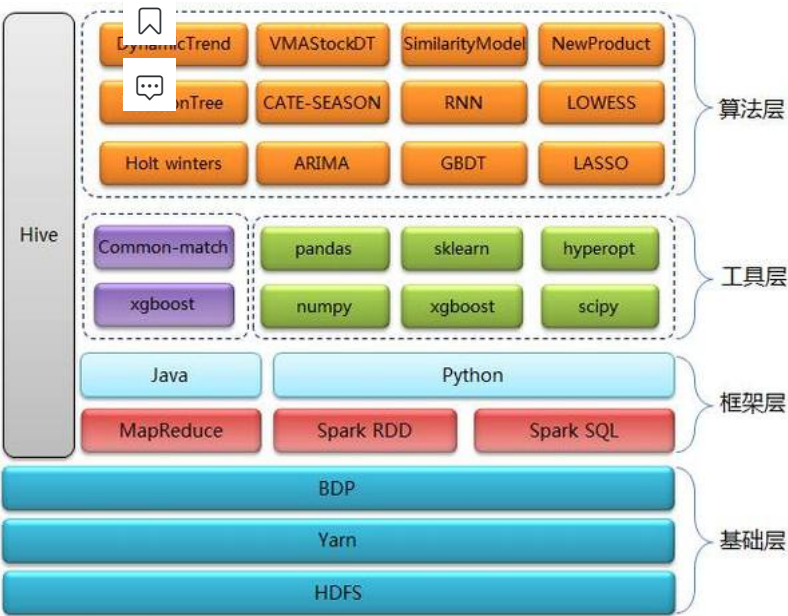
- 数据源输入层：京东数据仓库中存储着我们需要的大部分业务数据，例如订单信息、商品信息、库存信息等等。而对于促销计划数据则大部分来自于采购人员通过Web系统录入的信息。除此之外还有一小部分数据通过文本形式直接上传到HDFS中。
- 基础数据加工层：在这一层主要通过Hive对基础数据进行一些加工清洗，去掉不需要的字段，过滤不需要的维度并清洗有问题的数据。
- 核心业务层：这层是系统的核心部分，横向看又可分为三层：特征构建、预测算法和预测结果加工。纵向看是由多条业务线组成，彼此之间不发生任何交集。
  - 特征构建：将之前清洗过的基础数据通过进一步的转化化成标准格式的特征数据，提供给后续算法模型使用。
  - 核心算法：利用时间序列分析、机器学习等人工智能技术进行销量、单量的预测，是预测系统中最为核心的部分。



- 预测结果输出层：将最终预测结果同步回京东数据仓库、MySQL、HBase或制作成JSF接口供其他系统远程调用。
- 下游系统：包括下游任务流程、下游Web系统和其他系统。

### 3. 预测系统核心介绍

#### 3.1 预测系统核心层技术选型



预测系统核心层技术主要分为四层：基础层、框架层、工具层和算法层

**基础层：** HDFS用来做数据存储，Yarn用来做资源调度，BDP ( Big Data Platform ) 是京东自己研发的大数据平台，我们主要用它来做任务调度。

**框架层：** 以Spark RDD、Spark SQL、Hive为主， MapReduce程序占一小部分，是原先遗留下来的，目前正在逐步替换成Spark RDD。选择Spark除了对性能的考虑外，还考虑了Spark程序开发的高效率、多语言特性以及对机器学习算法的支持。在Spark开发语言上我们选择了Python，原因有以下三点：

- Python有很多不错的机器学习算法包可以使用，比起Spark的MLlib，算法的准确度更高。我们用GBDT做过对比，发现xgboost比MLlib里面提供的提升树模型预测准确度高出大概5%~10%。虽然直接使用Spark自带的机器学习框架会节省我们的开发成本，但预测准确度对于我们来说至关重要，每提升1%的准确度，就可能会带来成本的成倍降低。
- 我们的团队中包括开发工程师和算法工程师，对于算法工程师而言他们更擅长使用Python进行数据分析，使用Java或Scala会有不小的学习成本。
- 对比其他语言，我们发现使用Python的开发效率是最高的，并且对于一个新人，学习Python比学习其他语言更加容易。

**工具层：** 一方面我们会结合自身业务有针对性的开发一些算法，另一方面我们会直接使用业界比较成熟的算法和模型，这些算法都封装在第三方Python包中。我们比较常用的包有xgboost、numpy、pandas、sklearn、scipy和hyperopt等。

**Xgboost：** 它是Gradient Boosting Machine的一个C++实现，xgboost最大的特点在于，它能够自动利用CPU的多线程进行并行，同时在算法上加以改进提高了精度。

**numpy：** 是Python的一种开源的数值计算扩展。这种工具可用来存储和处理大型矩阵，比Python自身的嵌套列表结构要高效的多（该结构也可以用来表示矩阵）。

pandas：是基于NumPy的一种工具，该工具是为了解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。

sklearn：是Python重要的机器学习库，支持包括分类、回归、降维和聚类四大机器学习算法。还包含了特征提取、数据处理和模型评估三大模块。

scipy：是在NumPy库的基础上增加了众多的数学、科学以及工程计算中常用的库函数。例如线性代数、常微分方程数值求解、信号处理、图像处理和稀疏矩阵等等。

**算法层：**我们的算法模型非常多，原因是京东的商品品类齐全、业务复杂，需要根据不同的情况采用不同的算法模型。我们有一个独立的系统来为算法模型与商品之间建立匹配关系，有些比较复杂的预测业务还需要使用多个模型。我们使用的算法总体上可以分为三类：时间序列、机器学习和结合业务开发的一些独有的算法。

### 1. 机器学习算法主要包括GBDT、LASSO和RNN：

GBDT：是一种迭代的决策树算法，该算法由多棵决策树组成，所有树的结论累加起来做最终答案。我们用它来预测高销量，但历史规律不明显的商品。

RNN：这种网络的内部状态可以展示动态时序行为。不同于前馈神经网络的是，RNN可以利用它内部的记忆来处理任意时序的输入序列，这让它更容易处理如时序预测、语音识别等。

LASSO：该方法是一种压缩估计。它通过构造一个罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。用来预测低销量，历史数据平稳的商品效果较好。

### 2. 时间序列主要包括ARIMA和Holt winters：

ARIMA：全称为自回归积分滑动平均模型，于70年代初提出的一个著名时间序列预测方法，我们用它来主要预测类似库房单量这种平稳的序列。

Holt winters：又称三次指数平滑算法，也是一个经典的时间序列算法，我们用它来预测季节性和趋势都很明显的商品。

### 3. 结合业务开发的独有算法包括WMAStockDT、SimilarityModel和NewProduct等：

WMAStockDT：库存决策树模型，用来预测受库存状态影响较大的商品。

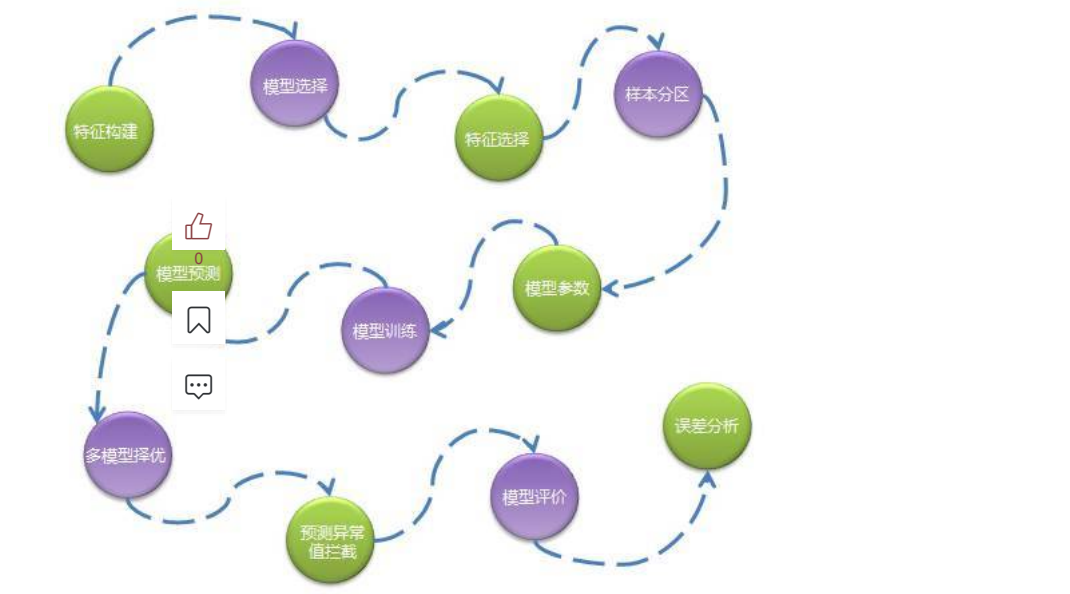
SimilarityModel：相似品模型，使用指定的同类品数据来预测某商品未来销量。

NewProduct：新品模型，顾名思义就是用来预测新品的销量。

## 3.2 预测系统核心流程

预测核心流程主要包括两类：以机器学习算法为主的流程和以时间序列分析为主的流程。

### 1. 以机器学习算法为主的流程如下：



**特征构建：**通过数据分析、模型试验确定主要特征，通过一系列任务生成标准格式的特征数据。

**模型选择：**不同的商品有不同的特性，所以首先会根据商品的销量高低、新品旧品、假节日敏感性等因素分配不同的算法模型。

**特征选择：**对一批特征进行筛选过滤不需要的特征，不同类型的商品特征不同。

**样本分区：**对训练数据进行分组，分成多组样本，真正训练时针对每组样本生成一个模型文件。一般是同类型商品被分成一组，比如按品类维度分组，这样做是考虑并行化以及模型的准确性。

**模型参数：**选择最优的模型参数，合适的参数将提高模型的准确度，因为需要对不同的参数组合分别进行模型训练和预测，所以这一步是非常耗费资源。

**模型训练：**待特征、模型、样本都确定好后就可以进行模型训练，训练往往会耗费很长时间，训练后会生成模型文件，存储在HDFS中。

**模型预测：**读取模型文件进行预测执行。

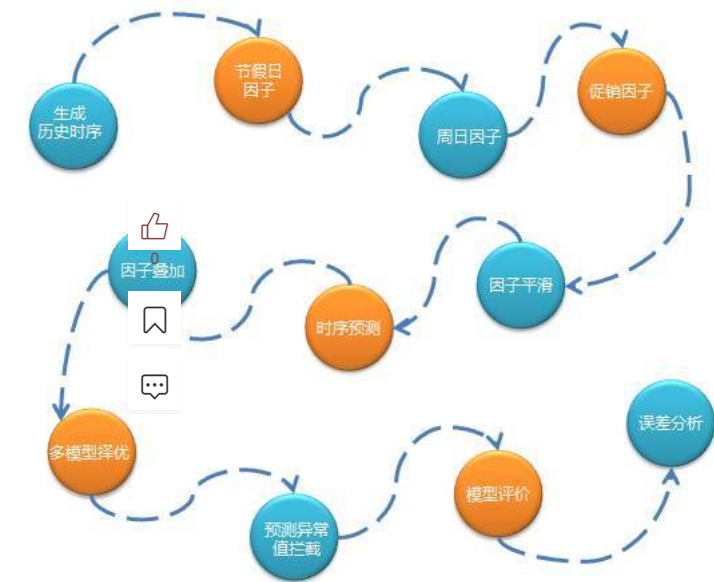
**多模型择优：**为了提高预测准确度，我们可能会使用多个算法模型，当每个模型的预测结果输出后系统会通过一些规则来选择一个最优的预测结果。

**预测值异常拦截：**我们发现越是复杂且不易解释的算法越容易出现极个别预测值异常偏高的情况，这种预测偏高无法结合历史数据进行解释，因此我们会通过一些规则将这些异常值拦截下来，并且用一个更加保守的数值代替。

**模型评价：**计算预测准确度，我们通常用使用mapd来作为评价指标。

**误差分析：**通过分析预测准确度得出一个误差在不同维度上的分布，以便给算法优化提供参考依据。

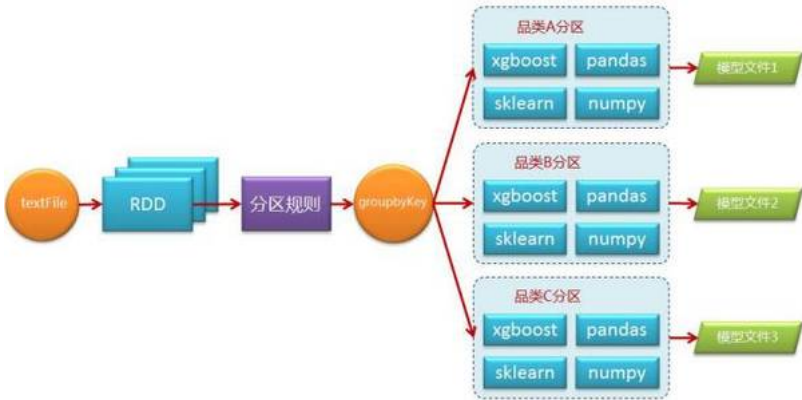
2. 以时间序列分析为主的预测流程如下：



### 3.3 Spark在预测核心层的应用

我们使用Spark SQL和Spark RDD相结合的方式来编写程序，对于一般的数据处理，我们使用Spark的方式与其他无异，但是对于模型训练、预测这些需要调用算法接口的逻辑就需要考虑一下并行化的问题了。我们平均一个训练任务在一天处理的数据量大约在500G左右，虽然数据规模不是特别的庞大，但是Python算法包提供的算法都是单进程执行。我们计算过，如果使用一台机器训练全部品类数据需要一个星期的时间，这是无法接受的，所以我们需要借助Spark这种分布式并行计算框架来将计算分摊到多个节点上实现并行化处理。

我们实现的方法很简单，首先需要在集群的每个节点上安装所需的全部Python包，然后在编写Spark程序时考虑通过某种规则将数据分区，比如按品类维度，通过groupByKey操作将数据重新分区，每一个分区是一个样本集合并进行独立的训练，以此达到并行化。流程如下图所示：






伪码如下：

```
sc.textFile("...").map(lambda x: repartitionBy(x)).groupByKey()\n.map(lambda x: train(x)).saveAsPickleFile("...")
```

repartitionBy方法即设置一个重分区的逻辑返回(K,V)结构RDD，train方法是训练数据，在train方法里面会调用Python算法包接口。saveAsPickleFile是Spark Python独有的一个Action操作，支持将RDD保存成序列化后的sequenceFile格式的文件，在序列化过程中会以10个一批的方式进行处理，保存模型文件非常适合。

虽然原理简单，但存在着一个难点，即以什么样的规则进行分区，key应该如何设置。为了解决这个问题我们需要考虑几个方面，第一就是哪些数据应该被聚合到一起进行训练，第二就是如何避免数据倾斜。


针对第一个问题我们做了如下几点考虑：

- 被分在一  区的数据要有一定的相似性，这样训练的效果才会更好，比如按品类分区就是个典型例子。
- 分析商品的特性，根据特性的不同选择不同的模型，例如高销商品和低销商品的预测模型是不一样的，即使是同一  使用的特征也可能不同，比如对促销敏感的商品就需要更多与促销相关特征，相同模型相同特征的商  倾向于分在一个分区中。

针对第二个问题  我们采用了如下的方式解决：

- 对于数据量过大的分区进行随机抽样选取。
- 对于数据量过大的分区还可以做二次拆分，比如图书小说这个品类数据量明显大于其他品类，于是就可以分析小说品类下的子品类数据量分布情况，并将子品类合并成新的几个分区。
- 对于数据量过小这种情况则需要考虑进行几个分区数据的合并处理。

总之对于后两种处理方式可以单独通过一个Spark任务定期运行，并将这种分区规则保存。

 目前您尚未登录，请 [登录](#) 或 [注册](#) 后参与评论

Spark-TimeSeries使用方法

 qq\_30232405 2017年04月24日 17:41  4372

1.spark里面的库是没有时间序列算法的，但是国外有人已经写好了相应的算法。其github网址是:<https://github.com/sryza/spark-timeseries> 但基本国内没...

时间序列分析--ARIMA模型

 u013527419 2016年10月15日 14:08  22615

<http://blog.csdn.net/u010414589/article/details/49622625> 指数平滑法对时间序列上连续的值之间的相关性没有要求。但是，如果你想使用指数平滑法计算...

程序员不会英语怎么行？

老司机教你一个数学公式秒懂天下英语

广告



基于spark用线性回归 ( linear regression)进行数据预测

ubuntu+spark+scala实现线性回归 ( linear regression ) 算法 ( 代码+数据 )

 wtt561111 2017年03月08日 13:05  3658

ARIMA模型

 u010159842 2017年06月07日 10:16  1355

ARIMA模型 自回归移动平均模型(Autoregressive Integrated Moving Average Model,简记ARIMA) 目录 [隐藏] 1 ...

实例讲解spark在京东智能供应链预测系统的应用

 woddle 2018年02月25日 18:08  166

问题导读：1. 京东的供应链是什么样的呢？2. 预测技术在京东的供应链起着什么样的作用呢？3. 京东整个预测系统的架构是什么样的呢？4. 预测系统不同层面的技术选型分别为什么？5...



多商城系统

开源多用户商城系统细节

百度广告



京东推荐系统：钱——打造千人千面的个性化推荐引擎

0

京东推荐产品及架构通用模型的应用离线CTR预测实例实验与监控京东推荐产品 80+推荐产品，包括移动端和web端 20+推荐服务，支撑EDM、微信端等 遍布用户网络的各个环节推荐系统的价值 挖掘用...



huizhejian

5年06月13日 17:54

3217



大数据|Spark技术在京东智能供应链预测的应用案例深度剖析（一）

大数据|Spark技术在京东智能供应链预测的应用案例深度剖析（一） 2017-03-27 11:58 浏览次数：148 1. 背景 前段时间京东公开了面向第二个十二年的战略规划，表...



javastart

2017年03月28日 19:58

1232

Spark技术在京东智能供应链预测的应用



javastart

2017年03月10日 19:05

630

Spark技术在京东智能供应链预测的应用 原创 2017-03-06 杨冬越 郭景瞻 大数据杂谈 大家晚上好，做一个简单的介绍：我叫郭景瞻，来自京东，著有《图解Spark：核心...

京东商城总架构师、基础平台负责人刘海锋：京东双11创新技术实践

【CSDN现场报道】2016 年 11 月 18 日- 20 日，由 CSDN 重磅打造的年终技术盛会 —— “2016 中国软件开发大会”（Software Developer Confere...



u012562943

2016年11月21日 10:06

2175

TOP100summit：【分享实录】京东1小时送达的诞生之路

京东1小时送达的诞生之路



msup789

2017年10月25日 16:33

396

码农怎能不懂英语？！试试这个数学公式

老司机教你一个数学公式秒懂天下英语



京东供应链溯源防伪平台



tigerking1017

2018年01月17日 15:51

388

PPT整理自：“别人在忙挖矿，京东架构师却悄悄用区块链搞了件大事！” [http://blog.csdn.net/dev\\_csdn/article/details/79062081](http://blog.csdn.net/dev_csdn/article/details/79062081) 仅供参考学习...

【智能制造】爱（AI）在新工业



np4rHI455vg29y2

2017年12月06日 00:00

275

本文是工业4.0俱乐部秘书长杜玉河老师在2017国际人工智能大会发表演讲时所用的PPT。人工智能是我们近期比较关注的领域，这篇PPT较全面地论述了人工智能的发展前景及场景应用。 ...

【揭秘：刘强东9年密谋的商业布局—京东快物流背后的核心技术盘点】

【揭秘：刘强东9年密谋的商业布局—京东快物流背后的核心技术盘点】 黄刚-物流与供应链 原创 2016-06-16 00:20:30 阅读数：4865 首次全面盘点刘强东9年物流布局，全面梳理...



javahongxi

2016年06月16日 01:42

2105

京东一元抢宝系统的数据底座架构优化



bestlove12345

2016年07月19日 16:03

1239

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

刘强东学习亚马逊：控制供应链 技术是最大障碍

dc\_726

2012年09月22日 20:35

8652

2月20日，刘强东一改往日的西装革履，一身灰色休闲帽衫、带着浓浓的“扎克伯格”范儿，出现在媒体面前。“老刘越来越硅谷风格了。”恰好他近段时间正是在美国学习，人们不免有了这样的窃窃私语。这也是他自关闭微...

多商城系统

开源多用户商城系统细节

百度广告



实战：供应链应用大数据

dongzhumao86

2015年03月27日 13:14

1858

摘要: 随着供应链变得越来越复杂，必须采用更好的工具来迅速高效地发挥数据的最大价值。供应链作为企业的核心网链，将彻底变革企业市场边界、业务组合、商业模式和运作模式等。 第三产业供应链协同应用市场进...

区块链在供应链领域的应用

chenhaifeng2016

2018年02月04日 13:18

1066

近年来，区块链作为一种新兴的应用模式被不同行业广泛应用。在包括金融、物联网、社会公益、供应链等领域中，出现了很多应用落地的探索和尝试。其中，供应链领域由于具有市场规模大，及多信任主体、多方协作等特点，...

大数据|Spark技术在京东智能供应链预测的应用案例深度剖析

duke370503

2017年04月19日 16:34

417

4. 结合图解Spark书进行应用与优化 《图解Spark：核心技术与案例实战》一书以Spark2.0版本为基础进行编写，系统介绍了Spark核心及其生态圈组件技术。其内容包括Spark生态圈、...

京东推荐系统

u014411730

2017年12月07日 09:19

1117

京东推荐系统 编辑 删除 在电商领域，推荐的价值在于挖掘用户潜在购买需求，缩短用户到商品的距离，提升用户的购物体验。京东推荐的演进史是绚丽多彩的。京东的推荐起步于2012年，当时的推荐产...

Spark 在金融领域的应用——日内走势预测

A3301

2016年11月19日 11:57

1299

作者：李涛涛 通联数据 1. 同花顺收费版之走势预测 2014年后半年开始，国内 A 股市场可谓是热火朝天啊，路上的人谈的都是股票。小弟虽然就职金融互联网公司，但之前从来没有买过股票，但每天听着别...

加入CSDN，享受更精准的内容推荐，与500万程序员共同成长！

登录注册

https://blog.csdn.net/woddle/article/details/7937048110/10