doi: 10.7690/bgzdh.2020.05.005

HBase 在飞行模拟训练数据存储中的应用分析

吕　游[1]，周晓光[1]，张家叶子[2]
（1. 海军航空大学教练机模拟训练中心，辽宁 葫芦岛 125001；
2. 中国人民解放军 92941 部队 45 分队，辽宁 葫芦岛 125001）

摘要：为提升飞行员飞行模拟训练质量，对飞行模拟训练数据化存储方式进行研究。介绍 HBase 分布式数据库原理、飞行模拟训练数据，对数据在 HBase 分布式数据库中的存储结构进行设计，提出一种基于 HBase 分布式数据库的存储方案，实现了数据在分布式系统中的入库程序，并通过实验对存储方案进行测试分析。测试结果表明：该存储方案有效地解决了飞行模拟训练数据存储的需求，并提高了数据的访问效率。

关键词：飞行模拟训练；HBase；数据存储
中图分类号：TP391　　**文献标志码：**A

Application Analysis of HBase in Flight Simulation Training Data Storage

LYU You[1], Zhou Xiaoguang[1], Zhang Jiayezi[2]
(1. *Flight Simulation Training Center*, *Navy Aviation University*, *Huludao* 125001, *China*;
2. *No*. 45 *Team*, *No*. 92941 *Unit of PLA*, *Huludao* 125001, *China*)

Abstract: In order to improve the quality of pilot flight simulation training, the data storage mode of flight simulation training was studied. This paper introduces the principle of HBase distributed database and flight simulation training data, designs the storage structure of data in HBase distributed database, puts forward a storage scheme based on HBase distributed database, realizes the storage procedure of data in distributed system, and tests and analyzes the storage scheme through experiments. The test results show that this storage scheme effectively solves the requirement of data storage for flight simulation training and improves the data access efficiency.

Keywords: flight simulation training; HBase; data storage

0　引言

飞行模拟器[1]是一种人在回路中的仿真系统，是尽可能逼真地复现空中飞行环境的飞行模拟设备。随着仿真技术的快速发展，飞行模拟器的模拟飞行程度越来越高，使模拟飞行训练环节在飞行人员的培养训练过程中越来越重要[2]。模拟飞行训练可以高效、安全地帮助飞行员了解飞行品质、培塑规范驾驶习惯、纠正孤僻飞行动作；因此，飞行模拟训练数据具有很高的应用分析价值。以往飞行模拟训练数据多以文件方式进行存储，导致数据的可扩展利用率很低，笔者将飞行模拟训练数据用数据库进行数据化存储，再根据其数据分析需求，提出一种利用 HBase 分布式数据库对数据进行列式存储的方案，将飞行模拟训练数据按参数重新组织，依据 HBase 数据库的存储特性，将一个参数的全部及其相关数据存入数据表中的一行，实现将数据分布式列式存储模式，满足对飞行模拟训练数据的使用需求，提高了数据的使用效率，并为联合分析大规模数据提供了技术支撑。

1　背景介绍

1.1　分布式数据库 HBase

HBase 是一个面向列、基于键值对存储结构的分布式数据库[3-4]。传统的关系型数据库定位一个数据需要数据表、行、列 3 个元素定位到单元格数据，在 HBase 中采用了五元组定位方法即数据表、行键、列族、列标识符、时间戳 5 个元素定位到单元格数据。HBase 中单元格是一个有时间版本的字节数组，时间版本是数据写入单元格时由 HBase 自动分配的时间戳，数据在写入前需要转换成字符串格式，数据存储序列严格按行键字典序排序。HBase 面向列的存储结构不同于传统的关系型数据库，其数据表中的列被分组成列族（column families），通过列标识符（qualifier）区分共同列族中的列；因此，用 column family: qualifier 表示数据表中的一列。

如图 1 所示，HBase 的存储结构采用主从节点

收稿日期：2020-01-09；**修回日期：**2020-02-23
作者简介：吕　游（1988—），男，黑龙江人，硕士，助理工程师，从事飞行仿真技术研究。E-mail: Lvyou88627@163.com。

的分布式模式：主节点 HMaster 只负责管理从节点 Client，并不存储数据，HRegionServer 作为存储节点负责数据的存储[5-6]；HBase 的一个数据表在建立时会分配一个 HRegion，当数据表中数据量增长到超过一个阈值时，HRegion 会横向拆分成 2 个相对较小的 HRegion，新拆分的 HRegion 会根据负载均衡策略重新分配到相应的 HRegionServer 节点上，HBase 最终的数据文件被上传至 Hadoop 分布式文件系统 HDFS 中。HBase 分布式数据库还可以动态地增加节点，使存储系统易扩展、更灵活[7-9]。

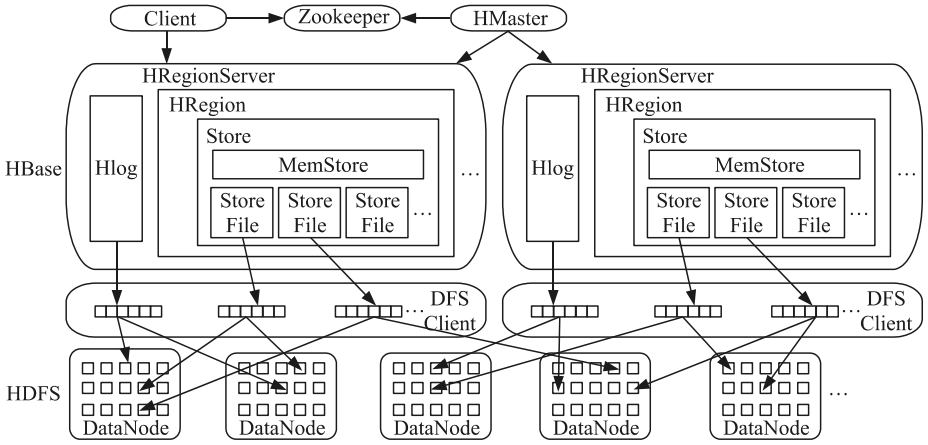

图 1 HBase 存储架构模型

1.2 飞行仿真数据

飞行模拟器通过仿真手段逼真地模拟飞机飞行过程，由周期性复杂仿真任务集构成的实时系统，其运行时系统数据流向如图 2 所示。整个模拟系统的输入端分为飞行员控制和教员指令 2 部分。系统首先接收飞行员控制杆的偏移信号，经过空气动力模型和运动方程解算出作用于飞机上的力和力矩；由运动方程模块综合气动力和力矩、发动机推力，并考虑燃油质量、飞行环境及教员指令的综合影响，解算出飞机的飞行状态、姿态和位置等数据，将解算出的数据通过外围通道传递给模拟器的其他分系统，飞行员再根据视景系统、仪表呈现的飞行状态进行下一步操纵。这样，一个数据周期就共同构成一个闭环的控制系统。

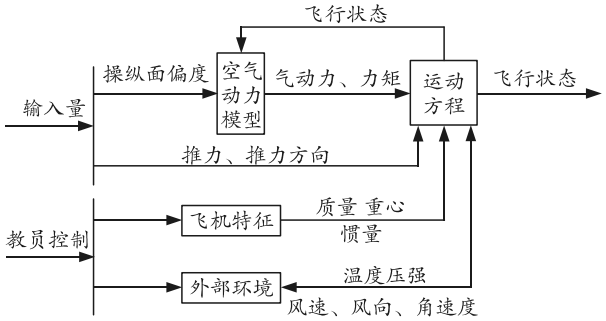

图 2 模拟器数据流向

飞行模拟训练过程产生的飞行模拟训练数据可以帮助分析相关型号飞机的飞行动力学模型设计是否合理，还可以辅助评估飞行员模拟飞行的质量和飞行结果的考核，并通过数据再次驱动视景仿真系统来重现飞行过程。飞行模拟系统将每个循环周期的模拟数据以行列表的结构写入文件中进行存储，一个固定的时间周期形成一个数据文件，数据文件包括文件头部描述信息和数据区 2 部分。头部描述信息包括飞行模拟器型号编号、记录数据的生成时间、参数名、数据类型等信息；数据区是一个行记录集，每行记录代表特定时刻所有参数的信号值，数据格式如图 3 所示。

```
12090;20181002;v1;v2;v3;v4;
21:05:18-30;21:05:18-262;21:05:18-39.936672;21:05:18-116.48437
21:05:18-266;21:05:18-2;21:05:18-39.885002;21:05:18-116.37381
21:05:18-265;21:05:18-294;21:05:18-39.01251;21:05:18-116.26835
21:05:18-194;21:05:18-185;21:05:18-39.37742;21:05:18-116.16346
21:05:18-126;21:05:18-225;21:05:18-39.336063;21:05:18-116.47849
21:05:18-296;21:05:18-147;21:05:18-39.583218;21:05:18-116.62151
21:05:18-143;21:05:18-175;21:05:18-39.95386;21:05:18-116.96616
21:05:18-0;21:05:18-236;21:05:18-39.88891;21:05:18-116.326126
21:05:18-255;21:05:18-256;21:05:18-39.033203;21:05:18-116.006615
21:05:18-51;21:05:18-151;21:05:18-39.744617;21:05:18-116.41195
21:05:18-315;21:05:18-289;21:05:18-39.998783;21:05:18-116.61961
21:05:18-95;21:05:18-21;21:05:18-39.547523;21:05:18-116.70313
21:05:18-301;21:05:18-169;21:05:18-39.409573;21:05:18-116.29854
21:05:18-53;21:05:18-126;21:05:18-39.975826;21:05:18-116.67392
21:05:18-143;21:05:18-90;21:05:18-39.026318;21:05:18-116.75373
21:05:18-307;21:05:18-272;21:05:18-39.903015;21:05:18-116.55884
21:05:18-237;21:05:18-23;21:05:18-39.84598;21:05:18-116.38504
21:05:18-111;21:05:18-62;21:05:18-39.76853;21:05:18-116.58155
21:05:18-43;21:05:18-154;21:05:18-39.570114;21:05:18-116.88375
21:05:18-305;21:05:18-75;21:05:18-39.324905;21:05:18-116.013245
21:05:18-329;21:05:18-5;21:05:18-39.991806;21:05:18-116.120384
21:05:18-342;21:05:18-178;21:05:18-39.482475;21:05:18-116.439445
```

图 3 飞行模拟训练数据文件

2 系统实现

2.1 存储结构设计

笔者介绍了 HBase 数据库存储结构不同于传统的关系型数据库，因其特有的存储结构，数据表结构设计对 HBase 数据库的应用效率至关重要[13]。HBase 数据库的主要优势在于列式存储特性。笔者设计将一个数据文件中的数据按参数重新组织划分，一个参数的所有数据存入到 HBase 单元格内，

与该参数的相应属性信息共同构成数据表中一行数据。HBase 数据库采用 Key-Value 形式存储数据，行键采用基于 Log-Structured Merge tree 的数据存储结构，数据存储是以行键为 Key 字典排序的，HRegion 是基于行键来为一个存储区域行数据集提供服务的。

一个含有 n 个参数、t 行数据的数据文件 f，可以描述为 $F^f=\left\langle H^f,RS^f\right\rangle$。式中：$H^f$ 为文件头部描述信息，包括飞行模拟器型号、数据生成时间、参数名等信息；RS^f 为数据文件 f 的数据区，包含 t 行数据，$RS^f=\left\langle R_1^f,R_1^f,\cdots,R_t^f\right\rangle$，其中 $R_j^f\ (1\leqslant j\leqslant t)$ 表示数据文件 f 的一行数据，$R_j^f=\left\langle c_{j1}^f,c_{j2}^f,\cdots,c_{jn}^f\right\rangle$，式中 c_{jk}^f 表示该行数据的 k 参数数据，这样数据文件 f 中参数 k 的全部数据就可以表示为 $c_{\bullet k}^f$。数据文件 f 转换到 HBase 数据表中生成 n 行数据，其中参数 k 的全部数据 $c_{\bullet k}^f$ 以字节序列的形式存储在数据表的一个单元格中，由于 HBase 列式数据库的特点，数据表可以自由增加数据列，根据参数 k 在未来的数据应用分析中的使用需求，对其全部数据 $c_{\bullet k}^f$ 进行预处理生成参数相关信息，如参数的极值、均值、方差等统计信息，也可以是参数的数据类型、取值区间、采集频率等属性信息，还可为简化仿真算法流程将数据经过高、低通滤波等处理结果信息[11-13]，用 CA_{ik}^f 来表示参数 k 的第 i 个关联信息单元格。至此，HBase 数据表的一行数据可以表示为 $HR_k^f=\left\langle RK_k^f,c_{\bullet k}^f,CA_{1k}^f,CA_{2k}^f,\cdots,CA_{sk}^f\right\rangle$，其中 RK_k^f 表示该行的行键，HBase 数据库中行键的设计尤为重要，直接影响了数据的检索效率，所以行键的设计需要结合飞行模拟训练数据分析检索需求和行键使用特点。文中行键 RK_k^f 设计由飞行模拟器型号、参数唯一编号和数据文件生成时间联合构成。这样的设计满足了按飞行模拟器型号、训练时间、特定参数的快速检索需求，能更好地发挥 HBase 数据库中 MapReduce 技术的长处，提高飞行模拟训练数据的访问效率。

2.2　数据入库过程

数据入库过程：首先将数据上传到分布式文件系统 HDFS，然后利用本地计算的优势通过 Map 过程将存储于本地的分布式文件系统的数据导入分布式数据库 HBase；数据在 HBase 集群入库时首先按 KeyValue 模式及笔者设计的表结构分解成行键 RowKey 和数据集，根据不同参数的关联属性需求调用函数处理数据集生成相应数据关联属性，与数据集一同构成列数据 Value。通过 Reduce 过程将数据写入到 Hbase 集群中，每一行数据的各项被存储在不同的列中，同列族的列集中存储在一个文件块中，HBase 数据库在写入数据的同时会自动检测 Rowkey 唯一性，如果冲突会增加数据版本并不删除数据。在此过程中，大部分的操作都在各个节点的本地进行，子节点和主节点之间只有少量的元数据信息交换；因此，数据导入时间很短。数据入库流程如图 4 所示，数据导入时主要调用 Importer 类实现并行入库，这个类继承自抽象 Mapper 类并实现了它的 map 函数。

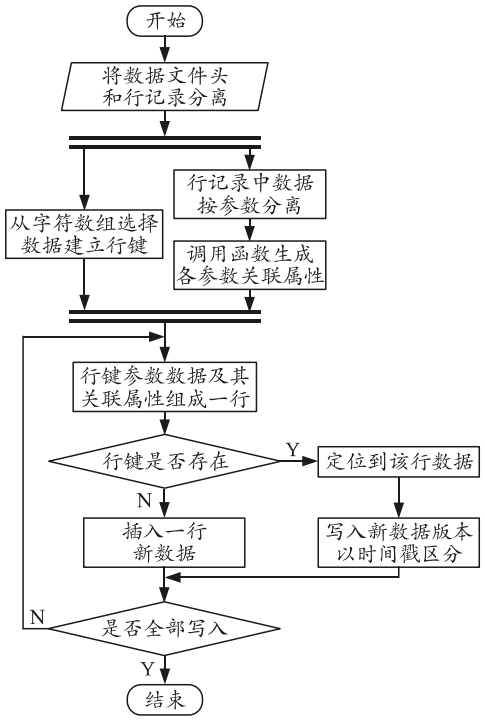

图 4　数据入库流程

3　分析与评估

3.1　实验环境

笔者搭建 3 台主机的分布式集群，操作系统采用 Ubuntu 11.04，设置 1 台主节点，2 台从节点。节点间网络带宽为 1 000 Mb/s。实验数据集选取某型飞行模拟器飞行训练 3 h 的连续时序数据记录，共 100 014 026 条记录，并存储在 10 000 个数据文件中。

3.2　实验结果

为验证笔者提出的基于 HBase 数据库存储方案

的可行性，首先对比相同实验条件下传统关系型数据库 MySQl 与 HBase 分布式数据库集群的数据入库效率；然后对比相同检索条件下，运用程序直接检索数据文件方法、关系型数据库 MySQl 和 HBase 分布式集群 3 种存储方案的检索效率。

笔者提出将飞行模拟训练数据进行数据化存储的方案，数据入库效率是影响数据存储系统整体效率的关键因素之一。如图 5 所示，实验记录了在不同数据规模下 3 种存储方案的入库时间，这里传统的以文件方式存储数据的入库时间记为文件拷贝时间。从图中可以看出：3 种方案数据入库时间均与数据规模呈正比，HBase 分布式集群发挥了 MapReduce 分布式算法的优势，数据入库效率优于关系型数据库 MySQL。

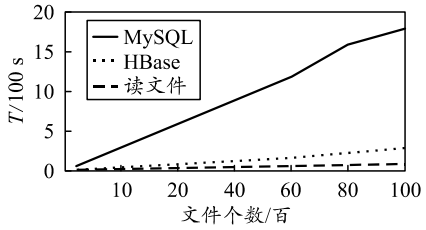

图 5 数据量与写入时间关系对比

在查询比对实验中选取查询指定参数的典型查询用例，执行 5 次后取平均值。如图 6 所示，实验设置了相同条件下 3 种方案获取某个参数全部数据的查询用例，将不同数据规模下的查询平均耗时记录绘图。从图中可以看出：3 种存储方案的查询时间均与数据规模呈正比，在该查询条件下 MySQL 关系型数据库和程序直接读取文件方式均需要扫描全部数据，由于直接读文件的存储方案可以借助文件的文件名检索，所以在该查询条件下，查询效率优于 MySQL 关系型数据库；由于 HBase 分布式集群的列式存储特性，以及其基于 Key-Value 形式的存储结构，所以其在该查询条件下效率优于其他 2 种方案，并且数据量在 2 000 个文件以内时，查询时间比较稳定，不随数据量增长而显著增长，在数据量超过 2 000 个文件时，由于数据分片及集群节点间通信影响，查询效率与数据量呈正比关系。

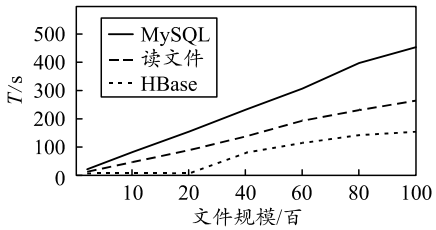

图 6 数据量与读取时间关系对比

4 结束语

笔者提出将分布式数据库 HBase 引入到飞行模拟训练数据存储中，针对数据应用特点和 HBase 数据库存储结构，设计了通过 HBase 数据库存储飞行模拟训练数据的存储表结构，并通过对比实验验证了在大数据量的情况下，该方案在数据入库和检索应用方面均优于传统的解决方案。下一步，笔者将 HBase 分布式数据库与关系型数据库联合使用，针对飞行模拟训练数据，引入智能评估体系。

参考文献：

[1] 李林. 飞行模拟器[M]. 北京: 北京理工大学出版社, 2012: 1-3.

[2] 林亚军, 王江南, 张原. 教练机飞行与指挥模拟训练系统建设[J]. 兵工自动化, 2019, 38(1): 17-20.

[3] KIM H J, KO E J, JEON Y H, et al. Migration from RDBMS to Column-Oriented NoSQL: Lessons Learned and Open Problems[M]. Proceedings of the 7th International Conference on Emerging Databases, 2018: 10-15.

[4] PALLAS F, GÜNTHER J, BERMBACH D. Pick your choice in HBase: Security or performance[C]. IEEE International Conference on Big Data, 2017.

[5] LARS GEORGE. HBase 权威指南[M]. 代志远, 刘佳, 蒋杰, 译. 北京: 人民邮电出版社, 2013: 303-304.

[6] PANDAGALE A A, SURVE A R. Hadoop-HBase for finding association rules using Apriori MapReduce algorithm[C]. IEEE International Conference on Recent Trends in Electronics, 2017.

[7] RAMANATHAN R, LATHA B. Towards optimal resource provisioning for Hadoop-MapReduce jobs using scale-out strategy and its performance analysis in private cloud environment[J]. Cluster Computing, 2018, 5(2): 1-11.

[8] 王永志, 包晓栋, 缪谨励, 等. 基于大数据的地质云监控平台建设与应用[J]. 地球物理学进展, 2018, 33(2): 850-859.

[9] BAO S, WEITENDORF F D, PLASSARD A J, et al. Theoretical and Empirical Comparison of Big Data Image Processing with Apache Hadoop and Sun Grid Engine[J]. Proc Spie Int Soc Opt Eng, 2017, 21(6): 10138.

[10] 张叶, 许国艳, 花青. 基于 HBase 的矢量空间数据存储与访问优化[J]. 计算机应用, 2015, 19(11): 3102-3105.

[11] 周胜明, 赵育良, 张玉叶, 等. 用于飞行动作评估的飞参数据预处理方法[J]. 兵工自动化, 2015, 34(5): 22-25.

[12] 胡文, 刘红军. 基于灰色理论的飞行模拟器逼真度评估[J]. 兵工自动化, 2007, 26(4): 18-20.

[13] 郭鹏程, 李迎春, 付春燕, 等. 海量日志数据采集系统的设计与优化[J]. 电子测量技术, 2018, 3(1): 12-17.