RubatoDB:一个新型 BigData 数据库管理系统

一、 大数据的挑战

当今，大数据已经不简简单单是数据大的事实了，而更重要的现实挑战是对多种类大容量数据存放，管理，分析，高效获取很多智能的，深入的，有价值的信息，这些信息在现实世界和智能虚拟领域为一个国家提供源源不断的发展动力。大数据已经是各个国家竞争的至高战略之一，美国如此，中国也如此。

大数据时代给原有的数据管理技术在高效，准确，安全方面都提出了前所未有的挑战。作为大数据管理的核心与应用基础，数据库管理系统的开发与使用是最重要的挑战之一。原先的数据库管理系统，已不能满足大数据应用所提出的需求。各个拥有绝对领先技术的国际专业公司（如 ORACLE，IBM，Google 等）都努力革新以适应大数据技术和应用发展趋势。数据库管理系统的技术已从原来单一数据库引擎发展到多引擎，从纵向扩展（SCALE IN）发展到横向扩展（SCALE OUT），从传统 SQL 发展到 NoSQL 数据库的提出和逐步实施等等。这些新技术为大量数据的集中存放和简单查询提供了解决方案。但是在目前市场上的数据库管理系统（无论是商用还是开源的）还没有一个能够同时支撑数据的高一致性，高可用性和高扩展性的新型数据库系统（NewSQL)。CAP 是新一代数据库管理系统的技术障碍和重要挑战。谁能抢先开发出新一代数据库管理系统，谁就能在这新的数据库管理系统的市场中抢先占有优势，就能在大数据战略发展中占有制高点。因此，具备原始创新和自主产权的新一代数据数据库管理系统则将为各国数据信息安全核心因素之一。

二、 NoSQL 系统

当传统 SQL 数据库管理系统在纵向扩展（SCALE IN）遇到了瓶颈时，NoSQL 数据库管理系统在 MapReduce (or Hadoop 平台）的分布运算模式的基础上做横向扩展（SCALE OUT），较好地解决了大数据的存储与简单查询问题。NoSQL 系统一个最重要的特性就是使用 X86 服务器集群以提供高可扩展性和高可用性（scalability and availability)。NoSQL 系统在大数据中得到广泛应用甚至被许多人称之为 NoSQL 革命，其重要性不可小觑。随着 NoSQL 系统的广泛应用，其各种弊端以及面临的几乎不可逾越的技术障碍也逐渐显露。

NoSQL 数据库系统的主要弊端是其低一致性。几乎所有 NoSQL 都依据 CAP 定律使用低一致性（BASE）而放弃了高一致性（ACID）。CAP 定律是建立在对分区容错性的侠义理解上的，但该定律实质上并不排斥 CAP 的三者兼顾。2014 年图林奖得主 Stonebraker 教授指出开发高一致性的数据库系统并不受 CAP 定律限制 [1]，并明确表明 NoSQL 系统所保证的低一致性

（BASE），放弃数据库系统的正确性是极不负责任的态度（Garbage）。

标准 SQL 的缺失是 NoSQL 数据库系统的另一弊端。以广泛使用的 Cassandra 为例， 其号称 SQL 简化版的 CQL 不仅不支持大数据分析中常用的集合函数和 GROUP BY, 连简单的非索引列（non-primary key or non indexed columns)的条件查询亦无法支持，尽管 Cassandra 在各网络公司得到广泛应用， 它很难用于各种复杂的大数据分析。

以 Oracle 和 DB2 为代表的经典数据库（Old SQL)以高一致性（ACID）和标准 SQL 的特点牢牢占据了 OLTP(事务应用）的几乎全部市场。但其高可扩展性的缺失和使用存储系统的高额代价使得经典数据库在大数据市场毫无用武之地。 借助其高可扩展性，NoSQL 系统得以广泛用于大数据的存储，但其低一致性使得难以用于任何需要事务处理的大数据系统，而标准 SQL 的缺失则进一步加重了应用系统的开发负担。

人们不禁要问，能否开发一套兼顾两者之长而避其短的大数据库系统呢？

三、 NewSQL 的兴起与挫折

自 2011 年起，国际学术和工业界逐渐认识到新一代的数据库系统必须兼顾经典 SQL 和 NoSQL 两者之长并能避其之短，并命名为 NewSQL 系统（新型数据库）。大家一致认为，NewSQL 系统必须具备以下特征 [2]：

a) 支持标准 SQL,
b) 保证高一致性（ACID）,
c) 高可扩展性（Scalability).

由于兼顾两者之长，NewSQL 系统既可以用于传统的事务处理系统（OLTP）亦可以用于各种大数据应用系统。NewSQL 的高扩展性不仅仅能带来高性能（high performance)和高可用性（availability)，亦摆脱了大型数据库应用系统对高价的操作系统和存储系统的依赖性（去 IOE）。这是因为集群中的任何一台服务器都可用于存储和各种数据库运算，因而无须诸如（IBM 或 EMC）SAN 的存储系统。

由于市场需要，各种 NewSQL 系统， 如 Google Spanner, VoltDB 等等, 应运而生。 但是，由于困惑学术界多年的分布式并发控制算法的缺失， 各种新出现的 NewSQL 系统并不支持高一致性。 比如， VoltDB 并不支持真正意义上的事务处理（Transaction），而是用 Stored procedures 来取代所有的事务。

尽管国际学术和工业界一致认为 New SQL 系统将会取代经典 SQL 和 NoSQL 系统而成为下一代的数据库系统，但近十年的系统研究与开发成果有限。

实际上大家都认识到 NewSQL 系统的真正兴起很大程度上取决于研发分布式并发控制算法的成功与否[3]。 加拿大阿尔伯塔大学计算机系终身教授袁立言潜心研究提出了以时间戳为基准的分布式并发控制算法并以此开发了 NewSQL 数据库系统 **RubatoDB**。其研究成果在国际数据库和大数据的一系列顶尖学术大会和学报发表 [3,4,5,6]，得到众多好评。譬如，A. Krechowics 等研究了近年来各种 New SQL 的理论研究和系统开发后，2021 年 5 月在 IEEE Access 发文高度评价基于时间戳的分布式控制算法，认为 RubatoDB 是目前最有前途的 New SQL 数据库系统之一（Rubato DB is one of the most promising systems) [7]。

四、 RubatoDB 的核心技术

加拿大阿尔伯塔大学计算机系终身教授袁立言，从事数据库理论和系统研究多年， 近年来研制了以时间戳为基准的分布式并发控制算法并以此开发了 NewSQL 数据库系统 **RubatoDB**。

RubatoDB 是满足大数据处理性能要求，提供标准 SQL(包括 ACID 事务处理)功能的 NewSQL 数据库管理系统。其主要特点如下

- 采用分布式并行计算架构，使用使用 X86 的服务器集群或云虚拟机集群，
- 支持标准 SQL 语言及通用 API, 包括 JDBC, ODBC, (.)NET，
- 保证高一致性（ACID）。
- 支持高一致性基础上的高可用性 （High Availabiity)。

RubatoDB 是国际上真正做到以上 4 点的极少数（其实是唯一）数据库系统。它采用了以下新技术：

- 基于 stage architecture（分级软件结构）的数据库系统结构。（注：MapRuduce 和 Hadoop 是简化的二级分级软件结构。）
- 使用自主创新的无锁公式并发控制算法。
- 使用自主创新的基于事务 LOG 的分布式冗余算法。

a) **数据库分级软件结构**

分级软件结构（stage architecture)系由美国 UC Berkeley 提出的用于 Web Server 的一种软件分级结构，其主要思想是大型服务性软件可以分割为一系列完全独立的程序模块。每个模块从前一个模块接受一条指令，用于执行一项简单功能，并将结果用指令形式发往下一模块。 RubatoDB 的数据库结构采用了分级软件结构，每一个 SQL 指令的处理有如接力比赛， 从数据库端口取得指令， 然后一级一级处理，最后一级将结果发往用户。使用分级软件结构的数据库主要长处是

- 软件开发简单，测试容易，可靠性高；

- 分级软件结构特别适合于分布式数据库系统：将各种模块分布至不同的服务器节点即可。

b) **无锁扣的公式并发控制算法**

无锁扣的公式并发控制算法其实是传统的多版本时间戳并发控制算法（MVCC with timestamps)的一种独特实现方法。系统在每一个数据表内设置一个微型内存数据库以存储当前 update 公式，并使用两级提交(two phase commit)的算法以达到分布式高一致性[4,5]。我们首先证明了这算法的正确性，并使用 TPC-C 标准测试结果成功表明我们的算法很好地解决了困惑学术界很久的分布式并发控制 难题[4]。

c) 新的基于事务 LOG 的分布式冗余算法

分布式冗余采取通常的主从组合（Master 和 Slave）。 为了保证分布式的高一致性， Slave 节点的同步系基于（1） Slave 节点的事务 LOG 和整个分布式数据库系统的全部已经提交成功的事务清单（List of all Committed Transactions since the last checkpoint)。

上述描述的 RubatoDB 核心技术已经获得国家发明专利。

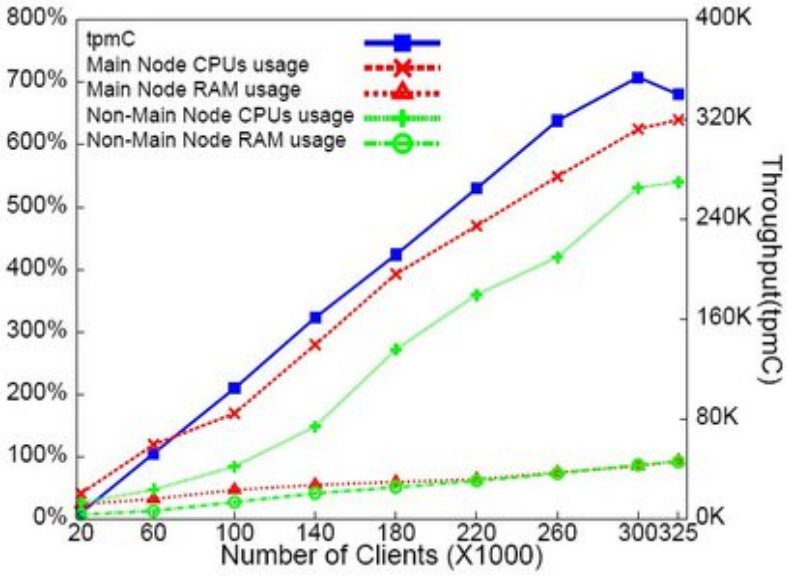
系统性能测试

我们使用了两种数据库标准测试结果以表明 RubatoDB 的结构和并发控制算法的优越性。

- TPC-C 标准测试:下表系 RubatoDB 的标准 TPC-C 测试结果.

	并发用户数	tpmC	每分钟事务处理数
1	25,000	28,935	65,761
4	85,000	105,572	240,618
8	160,000	184,524	419,372
16	325,000	363,759	828,725

- 使用 16 台(Linux) X86 服务器，存储 32，000 仓库（warehouse)的 TPC-C 数据库， RubatoDB 能够处理 32 万用户的并发需求。 下图显示出其性能指标线性增长的高扩展性 [5]。



- YCSB 大数据标准测试：RubatoDB 的大数据分析性能由下表的 YCSB 测试结果显示。（YCSB 是 YAHOO 提出 应用比较广泛的大数据测试标准）。

服务器数目	每秒运算数量（以读为主）	每秒运算数量（以写为主）
1	15,000	15,000
4	54,000	52,000
12	160,000	150,000

- 下图显示 RubatoDB 的大数据测试性能接近或超过主流 NoSQL 系统。

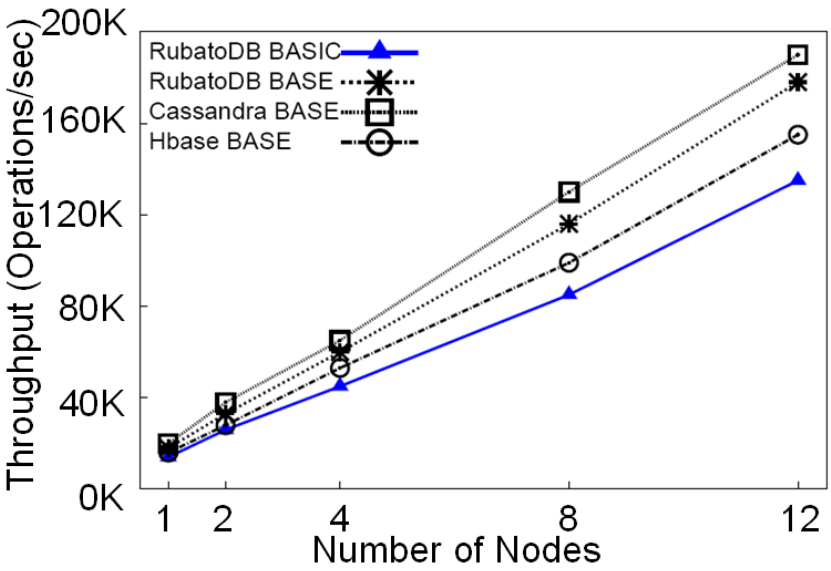


图 2：大数据 YCSB 测试

（注 BASIC 系我们提出的大数据系统的高一致性 [6])

d) **RubatoDB 的基本功能**

采用自主开发的核心技术和开源软件相结合的方式，我们已经基本完成了 RubatoDB 的商业软件系统开发。 其功能和前述的性能表明 RubatoDB 代表了新型数据库（NewSQL）系统。

- RubatoDB 在任意个（Linux）X86 服务器上作为一个整体数据库系统运行。（局限于实验室资源，目前测试的最高节点数为 64。）
  - 所有的服务器仅需网络连接（shared nothing)。
  - 每一个服务器皆可用于数据存储和 SQL 处理。
  - 支持动态节点扩展 / 收缩。
    - 超级用户可以任意在线增加新节点和退休（decommission)老节点。
- 支持标准 SQL 和各种标准界面
  - 支持创建各种 SQL 工具（schemata)，包括 table, view, index,trigger, stored procedure 等等。
  - 使用 MySQL 的通讯协议，因此用户可以使用各种 MySQL 的标准界面和工具以开发应用系统，包括
    - JDBC, ODBC, Python, PHP,
    - 各种 MySQL 的图形开发工具。
- 数据库表（SQL Table)可以任意分布到各个计算机节点
  - 建表语句支持下列三种 partitions

        - 节点分布（grid partition):每个计算机节点可以含任意个 grid partitions
        - 水平分布（hash (horizontal) partition)
        - 列分布（column partition） (currently beging re-coding)
    - 除了建表语句（Create Table Statement) 以及创建数据库语句（Create Database/Schema Statement)外，所有的分布都对用户透明
        - 所有 SQL 语句（除了建表／建数据库）都无须考虑分布状态。

- 支持标准 SQL 事务语句的最高一致性（ACID）

    - start transaction set isolation level serializable，
    - commit，
    - rollback。

  RubatoDB 是目前唯一支持 ACID 的分布式数据库系统。

- 支持高可用性（Availability)

    - 对于配置 N 个节点（组）的数据库系统， 任何一个节点（组）的故障并不影响系统其余 N-1 个节点（组）的正常工作。
    - 支持节点冗余（Replication)
        - 每一节点组可以配置 3 （或任意数目）的服务器（或虚拟机）以确保在 2 台服务器同时故障时的情况下，节点组仍能保持正常工作。
        - 采用基于事务 LOG 的分布式节点冗余算法（自主研发的世界领先的算法）以保证高一致性。
        - 每一个节点借助于系统生成的 Log 和 Snapshot 以确保节点及其数据可以自动恢复。恢复后的节点将以 Slave 的身份自动加入节点组。

- 支持 SQL Load 语句

    - 一个简单的 SQL Load 语句可以用来将数据文本（CSV file)上传至任何数据表
    - 数据上传速度达每个小时 4～8TB。实际上传速度取决于节点数目和网络速度

e) **RubatoDB 的主要优势**

- 信息安全

    - 国产数据库的知识产权在国内，是内资公司上海实方软件有限公司所拥有。这让信息安全有了基础。

- 实方软件公司目前正在进行基于 Application Level Encryption 的透明数据加密的开发。主要焦点是改进索引查询的速度。透明数据加密的开发成功将会进一步保障信息安全。

- 用于大数据和事务处理

  - 分布式架构使 RoubatoDB 能处理海量数据，能为大数据项目提供支撑处理大数据的能力
  - 高一致性使得 RoubatoDB 能用于关键数据的事务处理（OLTP）
  - RubatoDB 是国际国内真正做到即能处理大数据又能支持 ACID 强一致性的极少数（其实是唯一）数据库系统， 因此能为**去 IOE** 提供一个满意的解决方案。

- 系统保有成本低廉

  - RubatoDB 采用基于机架式 X86 服务器的架构，这让硬件架构简化了，不需要系统服务器＋存储阵列的架构
  - 系统维护人员技能要求的降低也降低了人员成本
  - RubatoDB 所需的服务器的数量可根据应用的需要逐渐扩展。项目初期可用少量服务器，随着数据量的增加，再逐步增加服务器。用户在初始的时候不需要大的投资

- 应用系统开发简便

  - 快速安装：Rubato 的半自动安装程序使数据库安装相当方便，可在短时间里完成大数据的数据库配置及安装。
  - 应用系统开发方便： 由于使用标准 SQL 及其各种标准界面，对开发员工的要求相对不高。即只需要懂 SQL 的工程师而不需要懂 Hadoop 等高级专家。这对团队的稳定有很大作用。（目前大数据数据库软件团队的稳定性的个较大的挑战）

- 大型关键系统（OLTP）的低成本灾变系统

  - 由于其标准 SQL 和高一致性， RubatoDB 可以作为大型关键系统（OLTP）的灾变系统
  - 由于使用机架式 X86 服务器的架构， 该灾变系统成本低廉
  - 除了 RubatoDB, 目前尚无任何合适的低成本灾变系统。

- 个性化的支持

  - 由于数据库核心是自主开发的，可根据应用程序的需求，为应用程序定制化的优化数据库的核心。这一点其它商业数据库是做不到的

f) RubatoDB 与其他各数据库系统的比较

下面我们将 RubatoDB 和其他各主要数据库作一简单比较

i. Oracle

- Oracle 作为经典 SQL 的代表系统， 功能强大，可靠性高， 品牌效益显著。 但系统无扩展性， 需要使用专用存储系统， 因此根本无法用于大数据系统。 系统保有成本极高。
- RubatoDB 使用 X86 服务器， 具有高扩展性，保有成本低。 无品牌效益。
- RubatoDB 是去 IOE 的潜在产品。

下表列出使用同样硬件成本的系统，RubatoDB 和 Oracle 的事务处理能力。

产品	Rubato DB	Oracle DB
测试标准	TPC-C	TPC-C
服务器成本	10万美元（2010价钱）	10万美元（2010价钱）
并发用户数	32.5万	22.3万
性能指标tpmC	36.3万	26万
数据来源	ACM CIKM2014	TPC官方网页

ii. Hadoop

- Hadoop 是一个基于 Java 的分布式系统的开发平台。 可以用于开发各种 NoSQL 系统。 能够提供高效高可用性的分布于数百甚至数千个 X86 服务器的大数据库系统。 但是 Hadoop 仅仅是一个开发平台， 使用 Hadoop 来开发大数据应用系统，需要高水准技术人员， 开发成本高， 而且系统性能指标往往差强人意。 当然，目前有各种基于 Hadoop 的系统可供选择。
- 除了不能支持标准 SQL, 几乎所有基于 Hadoop 的系统都无法提供高一致性。
- RubatoDB 是一个完整的 NewSQL 数据库系统， 无须再开发，因此应用系统开发简单方便。

下表显示出 RubatoDB 和 Oracle 以及 Hadoop 的主要差别。

	Oracle	Hadoop	Rubato DB
P级数据量	✗	✓	✓
线性扩展	✗	✓	✓
低硬件成本	✗	✓	✓
高性能	✗	✓	✓
高一致性	✓	✗	✓
SQL	✓	✗	✓
应用开发简单	✓	✗	✓
维护成本低	✓	✗	✓
品牌知名度	✓	✓	✗

iii. **MySQL**

- MySQL, 和 Oracle 类似， 系经典 SQL 的开源系统， 功能强， 可靠性高。 使用标准 SQL 界面， 应用开发简单方便。 但系统无扩展性，因此根本无法用于大数据系统。
- RubatoDB 使用 MySQL 的通讯协议和所有界面， 因此任何使用 MySQL 的应用系统可以直接转接到 RubatoDB 系统使用。 两者的主要差别是可扩展性（scalability)。 RubatoDB 可以视为“鸡血版的 MySQL“， 即一个存储容量， 并发用户数， 和系统吞吐量（system throughput)都能随意线性增长的 MySQL。

iv. **MySQL Cluster**

- MySQL Cluster 是 Oracle 基于单节点使用的 MySQL 而开发的分布式数据库系统的商业软件。 其最大优点是可扩展性，但不支持高一致性。
- MySQL Cluster 和 RubatoDB 的最主要差别是高一致性。

v. **Cassandra**

- Cassandra 是由 Facebook 牵头开发的开源 NoSQL 数据库系统，其主要优点是可扩展性和高可用性。Cassandra 使用的 CQL 查询功能极为有限，比如说， CQL 并不支持数据分析常用的集合函数（aggregate functions)和 Group By 功能，也不支持非索引列的查询条件。使用 SStable 的上传系统性能较差。使用 Cassandra 的系统进行数据分析非常困难。
- RubatoDB 的查询功能远远超过 Cassandra，而且其 SQL Load 的效率远远高于 Cassandra.

vi. **VoltDB**

- VoltDB （H-store 系其开源学术版）系由 Stonebraker 等知名教授开发的 NewSQL 数据库系统。其主要特征是（1） 高扩展性，（2） 内存数据库，以及（3）支持高一致性。公开发表的学术论文表明其在简化的 TPC-C 标准测试的性能（throughput)非常高。 该系统在学术界颇有定影响。 但该系统在分布式并非控制研发中没有实

质进展。实际上，该系统的 AICD 的达成仅仅靠（1）用 stored procedure 取代 SQL 事务，（2）任何两个分区（partition) 的冲突依靠锁住整个分区来保证。根据 TPC-C 测试中的远端事务(Guest warehouse)H-store 的吞吐量将急剧下降。

- RubatoDB 是目前唯一能提供有效分布式并发控制算法的 NewSQL 系统。根据标准 TPC-C 测试， RubatoDB 的并发用户数和吞吐量远高于 H-Store [4,5]。

下表列出 RubatoDB 和 VoltDB/H-Store 根据 TPC-C 标准测试的事务处理能力。 由于使用内存数据库和以存储过程（Stored procedure)取代 SQL 事务， 在 1%的远端事务中，H-Store 的吞吐量（throughput) 远高于 RubatoDB, 但随着远端事务增至 30%， 情况就完全改变了。 这完全得力于 RubatoDB 使用的自主开发的分布式并发控制算法。

产品	Rubato DB	H-Store	
测试标准	TPC-C	简化TPC-C （存储过程代事物）	
数据分区	16	16	64
并发用户数	32.5万	未公布	未公布
1% 远端仓库事务/秒	1.4 万	3.9 万	14.0万
30% 远端仓库事务/秒	0.7 万	0.7 万	1.8万
数据来源	ACM CIKM2014	ACM SIGMOD2012	

g) **Rubato DB 网址**

有关 RubatoDB 数据库管理系统的文件手册，有关论文及系统下载，均见以下网页：

- 中文：http://webdocs.cs.ualberta.ca/~yuan/databases/rubatodb/index.html
- 英文：http://www.rubatodb.com

五、 背景介绍

袁立言教授于 1981 年在上海交通大学获得硕士学位，1986 年在美国 Case Western Reserve University 获得博士学位。 其主持参与的基于其硕士论文提出的算法的某海军项目获得了中国国防科研三等奖。自 1986 年任职美国路易斯安那大学助理教授，1988 年任职加拿大阿尔伯塔大学计算机系，1992 年获终身副教授，1998 年获终身教授。他在国际顶尖数据库和大数据学术杂志上已发表了 100 多篇论文，并担任过国际数据库主要学术会议（VLDB，ACM PODS）的论文集编辑。他曾担任教育部“春辉计划”的特邀专家,”和长江学者”海外评审专家，并参与了科技部的 863 科技项目的研发。

附件：参考文献

[1] Michael Stonebraker, Errors in Database Systems, Eventual Consistency, and the CAP Theorem. BLOG@CACM, 2012.

[2] MichaelStonebraker, New SQL: An Alternative to NoSQL and Old SQL for New OLTP Apps, 2011.

[3] L. Wu, L.Y. Yuan, J.H. You. Survey and Taxonomy of Large-scale Data Management Systems for Big Data Applications JCST, 30(1), 2015。

[4] L.Y.Yuan, L. Wu, J.H. You, Y. Chi. Rubato DB: A Highly Scalable Staged Grid Database System for OLTP and Big Data Applications . ACM CIKM 2014.

[5] L.Y. Yuan, L. Wu, J.H. You, Y. Chi. A Demonstration of Rubato DB: A Highly Scalable NewSQL Database System for OLTP and Big Data Applications . ACM SIGMOD 2015.

[6] L. Wu, L.Y. Yuan, J.H. You. BASIC, an Alternative to BASE for Large-Scale Data Management System. IEEE Big Data 2014.

[7] A. Krechowicz, S. Deniziak, and G. Lukawski Highly Scalable Distributed Architecture for NoSQL Datastore Supporting Strong Consistency . IEEE Access, May 5, 2021.