企业级超融合数据平台

——用 ABCDE 赋能数字经济

2020 年 02 月

商标声明

longdb 和其他龙迪数智商标均为龙迪数智科技（北京）有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受龙迪数智科技商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，龙迪数智科技对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目录

第1章. 产品简介

在社交媒体、移动设备、应用场景激增的当下，业务数据的增长速度超过以往任何时代。为了增强企业竞争力，充分挖掘数据的价值，企业需要当机立断，引入新技术，辅助企业进行实时决策。实时决策要求在大数据技术上能够对数据量的增长和数据处理速度进行全面管理。

但当企业需要处理 TB 级到 PB 级的数据并即时做出决策时，当前的技术解决方案无法支持实时数据驱动的业务：

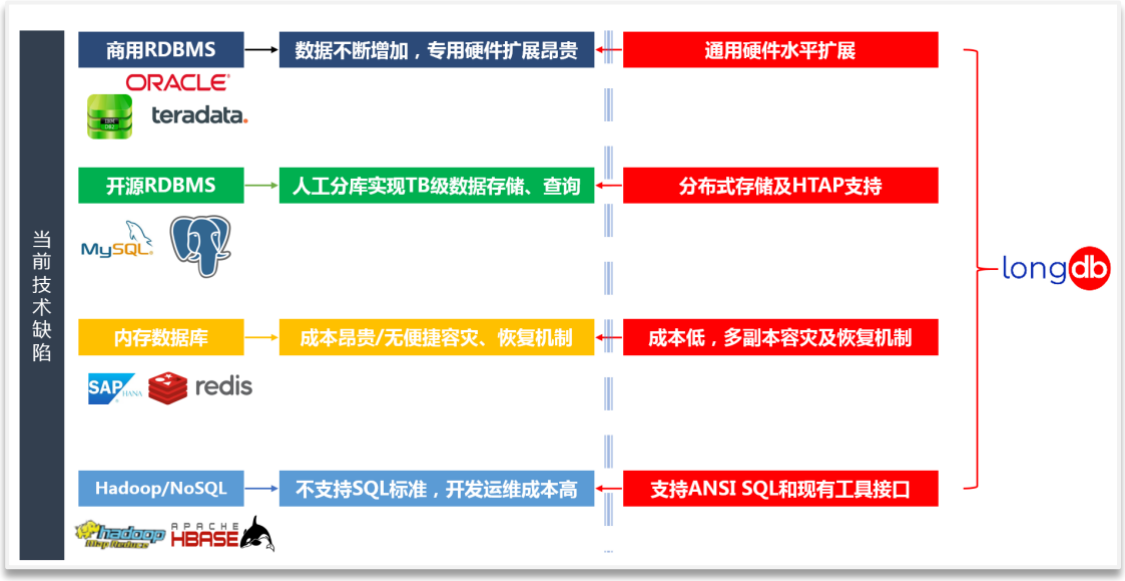


图 1

LongDB 以成熟的开源技术（如 Hadoop、HBase、Apache Spark）为基础，创建了一个灵活通用的数据库及大数据平台，图 1 展示了 LongDB 的主要优势：

- 速度快 – 大数据场景下，利用分布式 NoSQL 数据库 HBase 作为存储层，以及 Spark 的内存计算提供强大的数据处理能力；

- 低成本 – 满足 Hadoop 运行的通用硬件配置要求即可无缝横向扩展；
- ANSI SQL – 利用现有的基于 SQL 的分析，报表和应用无需重写；
- 分布式事务 – 基于 MVCC（多版本并发控制）的事务处理器，确保对多条记录和多个表的更新的同时数据的一致性；
- 灵活 – 以出色的性能同时兼容混合 OLAP（联机分析处理）和 OLTP（联机事物处理）工作负载；
- 弹性扩展 – 仅用几分钟便可扩展或收缩集群规模；
- 高频流式数据处理 – 支持高频的流式数据导入；
- 整合 Hadoop 生态 – 高度集成 Hadoop 开源社区组件，与 Spark 集成可实现 AI、ML 等应用场景。

以 LongDB 数据平台为核心基础，龙迪数智公司未来的产品发展愿景如图 2 所示：

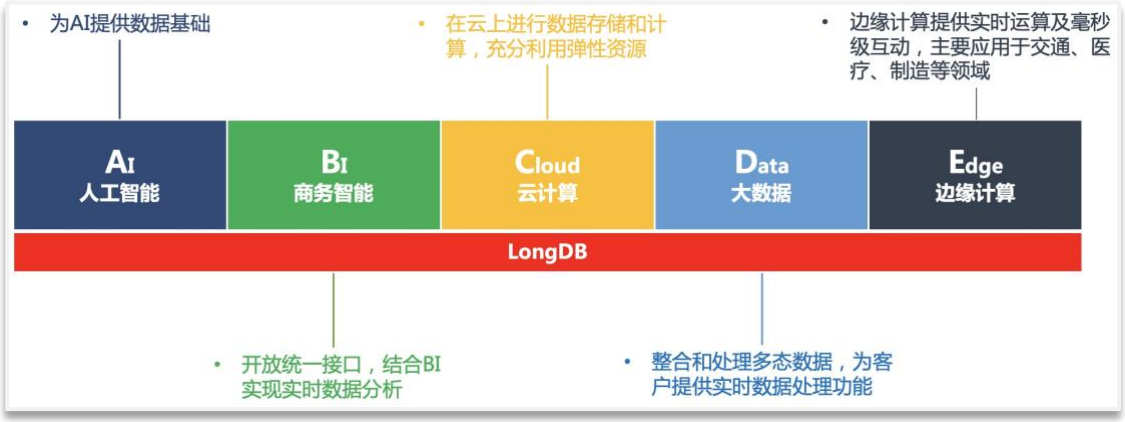


图 2

第2章. 体系架构

1. 稳定高效的分布式架构

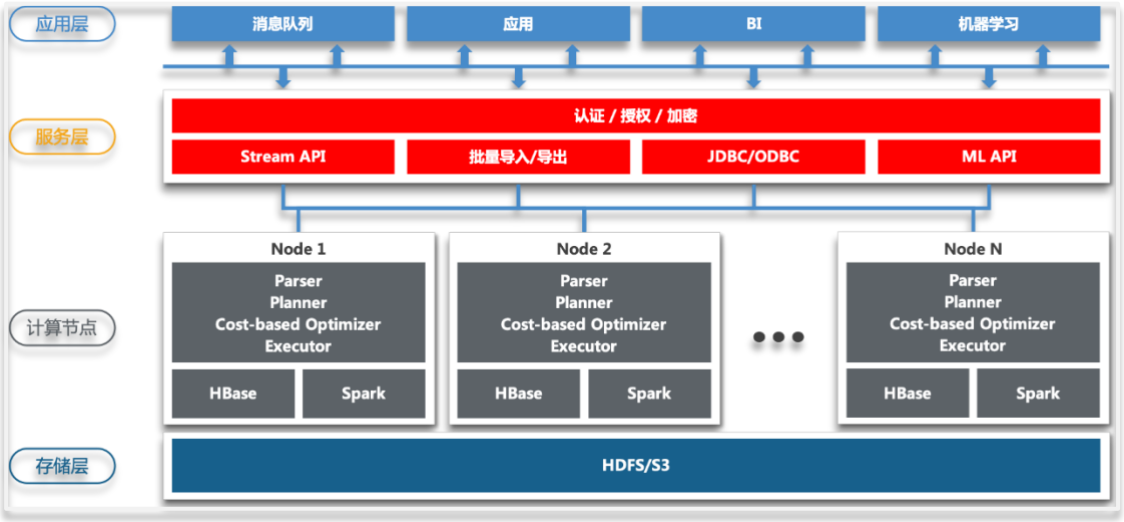


图 3

LongDB 实现了 Apache Kafka，Apache HBase 和 Apache Spark 的 Lambda 架构的功能，但避免了 Lambda 架构的复杂性以及数据的冗余，在同一个平台同时支持流式和批量数据处理。用户还可以选择使用流行编程语言或 BI 工具通过 JDBC/ODBC 等方式连接 LongDB，或者通过与 Spark 的专用接口无缝集成方式访问 LongDB，即可以直接使用 Scala、Python 和 R 语言操作 Spark DataFrame 中的结果集。

LongDB 可以从几个节点动态扩展到数千个节点，以支持各种规模的应用程序。

2. 高性能计算架构

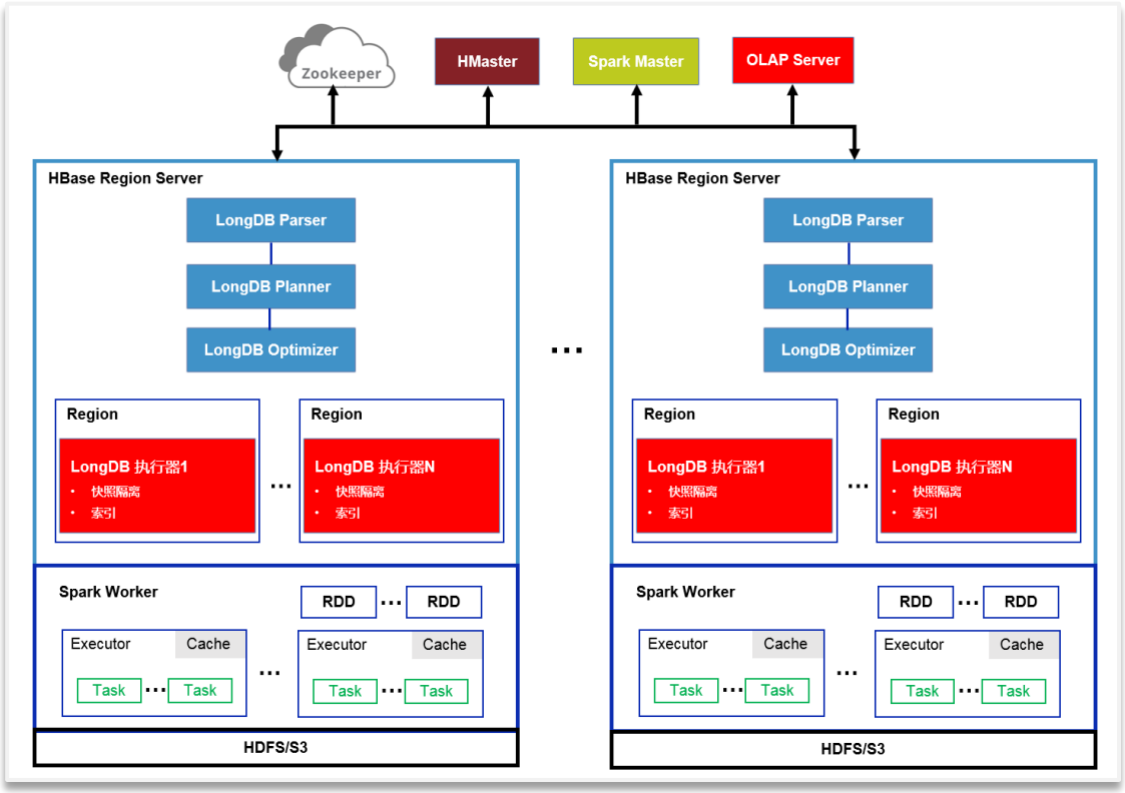


图 4

LongDB 拥有高性能的分布式计算架构，通过将计算推送到每个分布式数据分片，进行大规模并行化的谓词关联、聚合和函数运算。

在每个 HBase 物理节点上，LongDB 数据库为 HBase 和 Spark 提供了独立的进程和内存空间。它将解析器、规划器、优化器和执行器放在 HBase Region Server 进程中，并使用 HBase 协处理器跨区域（即分片）分发 OLTP（联机事物处理）计算任务。

3. 经济高效的混合引擎

图 5

LongDB 是基于 Spark 内存计算与基于 HBase 存储的 Hadoop 的混合 RDBMS（关系型数据库管理系统）。与内存数据库不同，LongDB 不会强制企业将所有数据都放在内存中，因为随着数据量的增长，这些数据存储成本会变得非常昂贵。LongDB 使用 Spark 内存计算来获得长时间运行查询的中间结果加速数据分析速度，而利用 HBase 的强大功能来持续存储和访问大规模数据。

当你在 LongDB 执行 SQL 时，它会根据集群上的数据分布并使用先进的基于成本的优化器来确定通过基表或索引对数据的最佳的访问方式、最佳的关联排序、最佳的分布式关联算法，以及它特有的根据查询特性和数据来选择执行查询的最佳计算引擎（例如 HBase 或 Spark）。单条记录查找/更新和小范围读取利用了 HBase 的高效检索功能，同时支持在 Spark 上长时间运行关联、聚合和分

组。

第3章. 产品特性

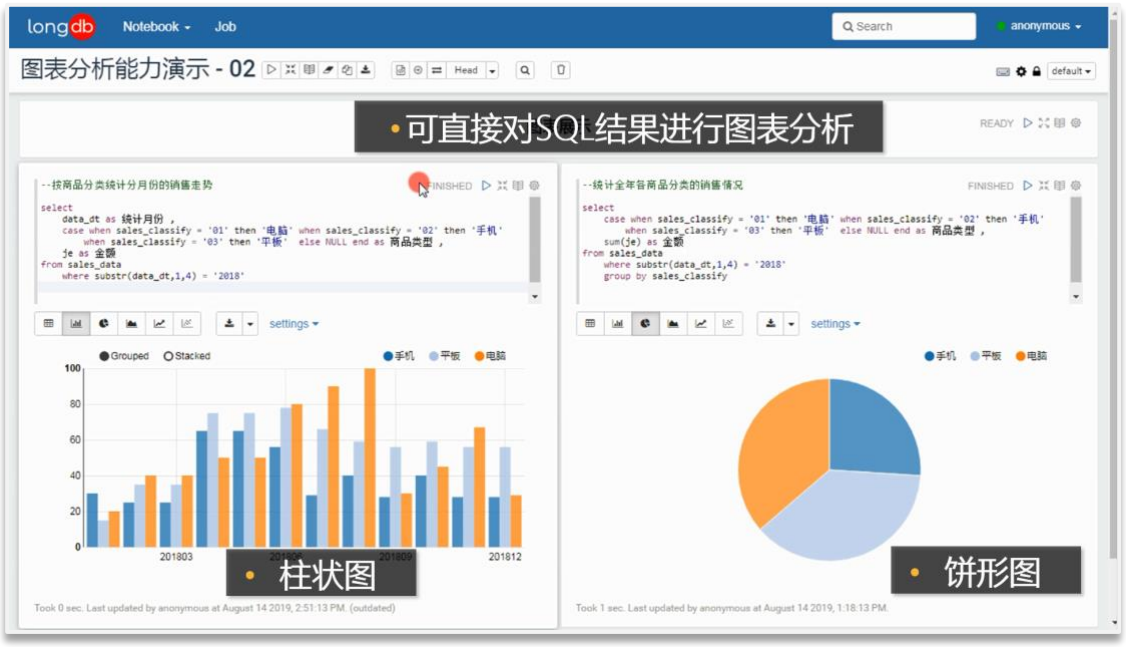
1. 简单易用的数据平台

LongDB 作为企业级数据平台，提供了方便快捷的部署工具，可以实现数据库的一键安装和部署，提供简单易用的开发和运维工具，极大的提高了工作效率。

系统管理平台：

LongDB 提供了基于 Web 的在线数据分析工具，多功能分析型 Notebook，可支持高达 20 种分析语言，可支撑数据集成、数据探索、数据分析、数据可视化等场景。

在线数据分析工具：



2. 支持标准 ANSI SQL

LongDB 拥有成熟的 SQL 处理技术，LongDB 在 Hadoop 生态提供了真正的关系型数据库，SQL 的兼容性达到了 ANSI SQL-2003 标准。

- Transactions
- Joins
- Secondary indexes
- Aggregations
- Sub-queries
- Triggers
- Constraints
- Column-level security
- Views
- Virtual tables
- External tables
- Window functions
- User-defined functions（UDFs）
- Stored procedures（Java/Python）

标准 SQL 使企业能够利用现有资源在 LongDB 上开发和维护应用程序。借助 ODBC 和 JDBC 驱动程序，现有应用可以轻松迁移到 LongDB，这些驱动程

序可以提供与 IBM Cognos®，SAP Business Objects®，Tableau® 和 MicroStrategy®等 BI 工具，Informatica®和 Ab Initio®等 ETL 工具，SAS® 和 R 等统计工具，以及 Toad®和 DbVisualizer®等 SQL 工具的无缝连接。

3．先进的基于代价的查询优化器

LongDB 以自研数据库核心为基础，并利用 HBase 作为它的存储层。LongDB 优化器能够利用收集到的数据统计信息并自动评估每个查询的代价将其发送到合适的计算引擎：OLTP（联机事物处理）查询（即小范围读/写或范围查询）转发到 HBase，OLAP（联机分析处理）查询（即大范围关联或聚合）转发到 Spark，计算引擎的切换自动完成。

4．利用高级内存技术加速查询

LongDB 集成了 Apache Spark（一种大规模数据处理的高速开源引擎）用于加速 OLAP（联机分析处理）查询。Spark 具有非常高效的内存处理机制，而且如果查询处理超出可用内存，它可以溢出到磁盘（而不是丢弃查询）。

更重要的是，Spark 对可能发生在集群中的节点故障具有优秀的控制能力，其他内存技术将丢弃与故障节点有关的所有查询，而 Spark 通过血缘方式在另一个节点上重新生成其内存中的弹性分布式数据集（RDD）。

5．利用商用硬件进行经济高效的扩展

LongDB 利用成熟的 Hadoop 分布式框架和通用服务器进行扩展。在像

Facebook 和阿里巴巴这样的企业，HBase 利用价格低廉的商用硬件，其集群支持的数据量已经扩展到几十至上百 PB。集群的节点数也已经上万台。

6. 同时支持 HTAP（OLTP/OLAP）混合工作负载

LongDB 数据平台的混合型架构旨在同时支持高性能的 OLTP（联机事物处理）和 OLAP（联机分析处理）查询，将 Spark 和 HBase 进行了无缝整合。LongDB 内置优化器自动评估每个查询并将其发送到适当的计算引擎，Spark 为 OLAP（联机分析处理）查询提供内存计算性能，而 HBase 可以用于扩展到 PB 级的 OLTP 查询。

资源隔离管理：LongDB 使用高级资源管理来确保同时进行 OLTP（联机事物处理）和 OLAP（联机分析处理）查询的高性能。通过对 Hadoop 和 Spark 的独立进程和资源的管理，LongDB RDBMS 可以确保复杂的 OLAP（联机分析处理）查询不会干扰对时间敏感的 OLTP（联机事物处理）查询。 用户可以为 OLAP（联机分析处理）查询设置自定义优先级，以确保在消耗所有群集资源的大量批处理过程中不会阻碍重要报表的生成。

7. 灵活通用的数据平台

LongDB 是一个灵活的数据平台，可以为 OLAP（联机分析处理）和 OLTP（联机事物处理）工作负载提供支持，LongDB 的存储层是 HBase，它是一个无模式的键值存储。 这为 LongDB 提供了灵活的存储数据和更新模式。

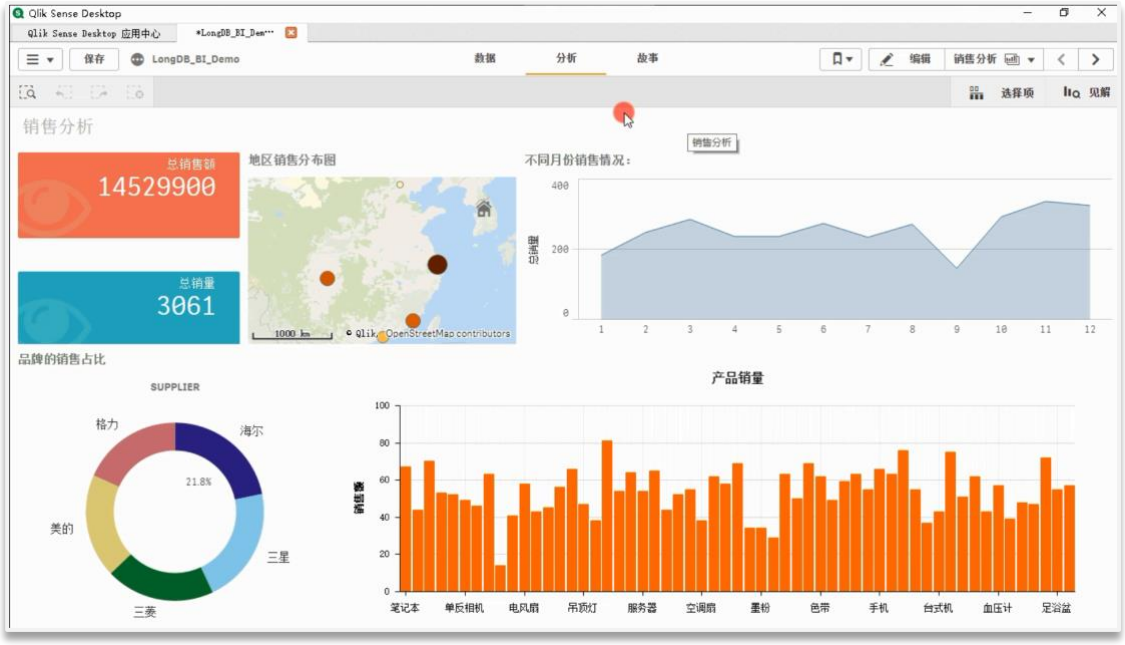
支持多种文件格式 LongDB 支持多种文件格式的导入、导出，包含但不限

于文本文件（例如 CSV、TXT 等）、Avro、列式存储（例如 ORC、Parquet 等）、压缩文件（例如 Snappy、GZ 等）、Spark 应用数据和流式数据。

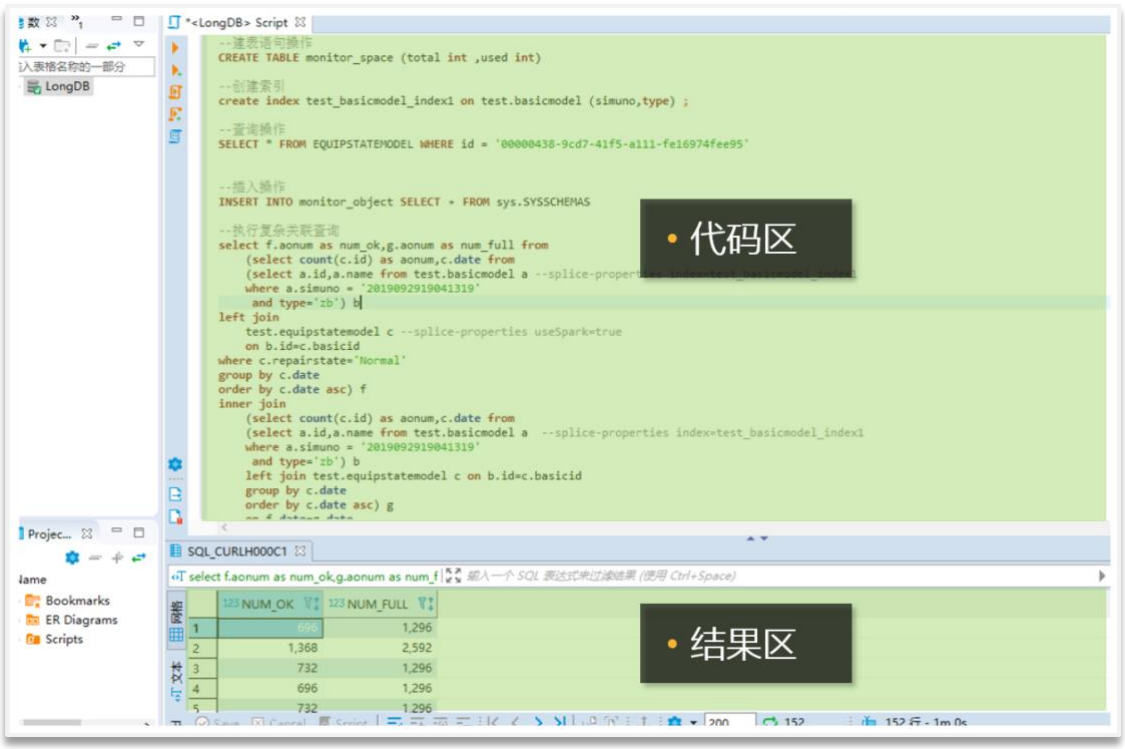
MapReduce 输入/输出格式 并行的 LongDB 查询利用 HBase 协处理器而不是 MapReduce 对存储在 Hadoop 分布式文件系统（HDFS）中的数据进行分布式计算。但是 LongDB 确实提供了一个 API 来针对存储在 LongDB 中的数据运行 MapReduce、Hive 和 Spark 作业。 这使用户能够使用 Storm、Kafka、Pig、MLlib 或 Mahout 等其他技术在 LongDB 中对整个数据集执行自定义的，面向批处理的分析。

支持 ODBC/JDBC LongDB 还提供了许多选项来访问其数据和跨服务器进行部署。应用程序和分析人员可以通过命令行界面（CLI）以及 JDBC / ODBC 连接使用 SQL。因此，开发人员可以使用各种编程语言（包括 Java、Scala、Python、C ++和 JavaScript 等）将 SQL 嵌入其代码中。

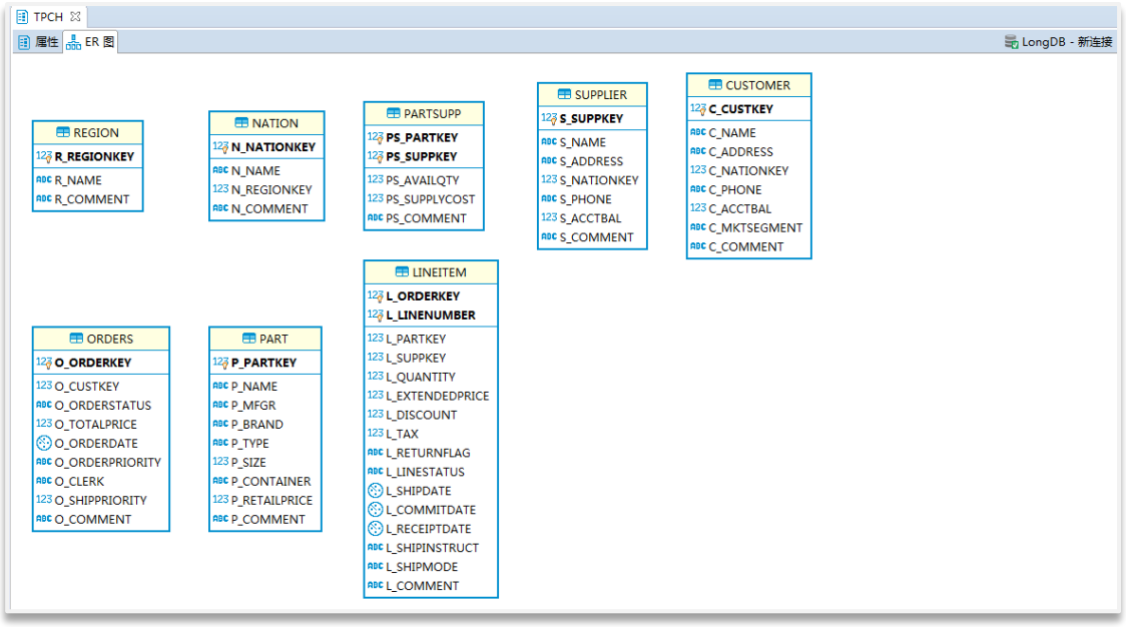
BI 展现工具可通过 JDBC/ODBC 接口来连接 LongDB 实现复杂的实时可视化分析，例如使用 Qlik Sense：



此外还可以使用第三方的数据库客户端管理工具对 LongDB 进行使用和管理，例如使用 DBeaver：



使用第三方数据库客户端查看数据库的 ER 图：



外部表 LongDB RDBMS 支持访问外部数据和库，可以使用虚拟表接口（VTI）对外部数据库和文件中的数据执行联合查询。它还可以执行所有预编译过的 Spark 库（超过 130 个并且不断增长），包括机器学习、流式分析、数据集成和图形建模。LongDB RDBMS 还支持外部表功能，可以在 LongDB 中建立外部表访问其他大数据组件中的数据（例如 Hive），并支持 ORC、Parquet 格式的外部表，同时其他大数据组件（例如 Hive）也可以创建外部表的方式访问位于 LongDB 上的数据，此特性提供了灵活的数据联邦查询功能。

兼容主流 Hadoop 发行版 随着 LongDB 在每个 HBase Region Server 上快速安装，运营团队可以继续使用他们现有的 Hadoop 发行版（例如 CDH、HDP、MapR）。通过 HBase 与 YARN 的集成，LongDB 可以与其他工作负载共享 Hadoop 集群，高度集成 Hadoop 开源社区组件，其中与 Spark 的集成可实现 AI、ML、流式数据等各种应用场景。

第4章. 应用场景

像 LongDB 这样的通用数据库几乎可以覆盖到每一个行业的诸多不同类型的实时应用。LongDB 的独特之处在于它能够实现新一代的实时分析应用，这些应用可以实时采集，分析和响应数据以改变行业现状。本节的其余部分重点介绍了 LongDB 可以支持的应用案例。

1. ETL 加速

ETL 处理是大多数 IT 部门的隐性负担，对于大数据而言，ETL 流程已成为依赖数据的应用和分析师的瓶颈。通过使用 LongDB RDBMS 替换像 Oracle RDBMS 一样不堪重负的操作型数据库，公司可以将 ETL 延迟从几天和几小时缩短到分钟级和秒级。LongDB RDBMS 整合了全世界最优秀的技术——RDBMS 的事务完整性和成熟的 Hadoop 横向扩展和 Spark 的内存计算性能。因此 LongDB 提供了一种更轻松的方法来加强 ETL 处理能力。

此外，在大数据处理场景下需要高速加载批量数据和流式数据，LongDB 通过其先进的技术，提供 Bulk Load 的方式保证数据的高速批量加载，提供 Spark Adaptor 实现流式数据的高并发高效率的计算和加载。

2. 替换旧有数据仓库

企业级数据仓库市场巨大，曾被类似 Teradata、Oracle、IBM 等巨头垄断，这些公司通常使用自有的分析型数据库系统承担数据仓库建设工作。随着数据仓

库客户业务的发展，企业数据仓库系统开始面临数据量膨胀速度快，数据分析类型更复杂、企业对数据响应速度高的压力。这迫使分析型数据库系统具备更强的存储和计算能力，才能满足企业客户数据分析需求的演进，扩展数据库硬件的成本和运维费用也比较高昂，加之 Hadoop 快速成熟、开源社区活跃，不满足需求的分析型数据系统开始逐渐被市场淘汰。而 LongDB 提供了功能完善的 SQL 在这样的背景下，LongDB 作为 Hadoop 上的新型数据库替换旧有数据仓库系统成为了最佳的选择。

3. 操作型数据湖

操作型数据湖（ODL）是操作型数据存储（ODS）的新型替代品。ODL 允许从更昂贵的 OLTP 和数据仓库系统支撑实时报表和分析，以及为 ETL 过程执行聚合和转换操作。

与使用纵向扩展技术的传统 ODS 相比，ODL 通过横向扩展技术提供 20 倍的性价比优势。它还可以处理半结构化和非结构化数据，作为更大的基于 Hadoop 的数据湖的一部分。

4. 数字营销

消费者营销优化了跨网站、移动设备、电子邮件和广告与数百万消费者间的互动。其成功的关键是实时个性化，在正确的时间向正确的人显示“正确的信息”。传统解决方案仅使用基于前一日数据的分析模型。LongDB 能够利用实时数据，而非其他系统延迟的 ETL 数据，帮助企业当机立断，实时决策。

5. 物联网应用

遥测是用于检测需要实时触发条件设备的数据流。电信、传统公用事业、互联网服务提供商（ISP）和电视有线/卫星公司等系统密集型行业使用实时遥测技术主动检测故障、隔离故障，然后尝试远程重置，以避免服务电话和派遣现场技术人员。

行业案例包括服务器监控，检测网站任何可能的性能下降，随后向受影响的用户提供优惠以提高客户留存率。网络安全应用监视防火墙，将实时活动与历史防火墙日志相关联，确定是否存在真正的威胁。 传统系统无法收集和分析庞大数量的系统遥测数据。LongDB 具有可扩展性、可以采集、分析和响应大量的遥测数据。

6. 精准医疗

随着医疗保健成本的持续上升，其增速已超过通货膨胀，美国等国家已经开始从基于活动的支付转向基于结果的支付。

医疗行业已开始关注精准医疗，该医疗使用基因学，临床和诊断数据来为特定个体量身定制治疗方案。基因数据可能很大，可以帮助预测结果和可能的并发症，以及为癌症等复杂疾病量身定制药物。临床数据包括电子健康记录（EHR），以确保在多个治疗和医疗提供者之间协调护理；诊断数据，尤其是来自医院和家中的实时设备数据，可以在并发症达到临界点之前触发干预，要求重新入院。

传统系统无法处理精准医疗所需的大量数据。LongDB 可以通过处理大量基

因组数据，为 EHR 提供数据存储计算以及收集大量实时设备数据来支持精准医疗。

7．运营管理支持

传统的 RDBMS 通常只支持实时交易系统(OLTP)，包括企业软件以及 Web、移动和社交应用，这些系统需要大量的事务完整性实时读写。通过 ACID（原子性、一致性、隔离性、持久性）事务，LongDB 可以为这些应用提供支持，同时通过商用硬件提供经济高效的扩展。

传统 RDBMS 上的操作分析系统（OLAP）传统上支持对事务和聚合数据的实时即席查询。然而，随着实时 Web、社交、移动和机器生成数据的爆炸式增长，传统系统无法满足扩展需求。

Teradata 和 Netezza 等专用设备可以扩展以处理海量数据集，但它们价格昂贵且需要大量的 ETL。LongDB 提供两全其美的优势：大规模数据的实时运营分析。

第5章. 特点总结

作为由 Hadoop 和 Spark 提供支持的**高性能融合型 RDBMS**，LongDB 为应用程序开发人员和数据库架构师提供了无限的可能性。最重要的是，它消除了迄今为止任何数据平台选择的局限性。您可以通过成熟的商用硬件进行**横向扩展**，内存优化技术可支持最大限度地发挥用户对标准 SQL 的操作和应用性能。

您可以降低购买专用服务器软硬件费用，您将拥有一个可**大规模扩展**并同时处理**操作型和分析型混合工作负载**的通用数据库。LongDB 还可以帮助您发掘当前 Hadoop 平台中已有数据的价值。

通过采用创新的混合架构，LongDB 具备独特的能力，为数据驱动型企业提供强有力的支持，使其可以利用实时洞察力来采取更好的行动。企业可以使用 LongDB 当机立断，实时决策，并超越竞争对手。**实现利用 ABCDE(AI, BI, Cloud, Data, Edge computing)技术最大化数字经济价值愿景。**