

# Selva 大数据平台

传统的工厂数据存储方式是采用实时数据库（包括实时部分和历史部分），该方式能很好的处理了传统的数据类型，例如：整形，浮点型，字符串等。但，目前工厂中的数据开始变得多样化。传统的数据类型以外，逐渐开始出现文本，图像，视频，声音等大容量的数据。这些数据都不能直接存储到实时数据库中，导致数据的关联查询越加困难。

同时，大数据的发展孕育出了高性能的开源非关系型数据库，以前在实时数据库上才能达到的数据处理能力，在大数据平台上已经能够触及，再加上大数据平台的灵活数据类型支持，动态处理节点的增减，更是优于目前的实时数据库。特别是大数据平台更适合数据挖掘。

关系数据库面临下列挑战：

并不适用于数据类型多样化（比如图像、视频和文本）的大量（PB 级）数据。

无法扩展、支持庞大的数据量。

无法纵向扩展，受制于内存和处理器的功能。

无法横向扩展，受制于依赖缓存的读取和写入操作。

分片（将数据库分成几个部分，存储在不同的节点）引起操作问题（比如管理共享式故障）。

复杂的 RDBMS 模型

一致性限制了 RDBMS 的可扩展性。

相比关系数据库，NoSQL 更具有可扩展性，并提供了出色的性能。NoSQL 数据库提供了下列解决方案，从而克服了关系模型无法克服的挑战：

横向扩展、什么都不共享的架构，能够在大量节点上运行。

非锁定并发性控制机制，那样实时读取不会与写入产生冲突。

可扩展的复制和分发一成千上万个机器拥有分布式数据。

每个节点提供的性能高于 RDBMS 的架构。

无模式 (schema-less) 数据模型。

### **本数据平台主要特点**

#### (1) 去中心化

数据库集群中的每一个节点的角色完全相同，避免了整个系统的单点问题导致的不稳定性。数据被分布在集群中的各个节点，每个节点包含不同的数据，但是没有主辅之分，每一个节点都可以响应所有请求。

#### (2) 支持复制和多数据中心之间的复制

提供可配置的复制策略，工厂大数据平台是作为分布式系统设计的，可以部署到包含多个数据中心的大量节点上。工厂大数据平台的分布式架构就是为多数数据中心部署、数据冗余、容错、灾难恢复等功能量身定制。

#### (3) 可扩展性

当数据库集群中增加新的节点时，数据库的读写吞吐量随之线性增长。并且不需要停止系统，不影响系统的运行。

#### (4) 容错性

数据可自动的复制到多个节点来实现容错性,支持跨越多个数据中心的复制。出错的节点可在不停止系统的情况下进行更替。

#### (5) 可调节的一致性

数据库的读写提供可调节的一致性水平,如所有节点的数据必须一致或者任一节点的数据都可使用,又或者要求一半以上的数据一致等。

#### (6) 支持 MapReduce

#### (7) 查询语言

工厂大数据平台引入了 CQL 查询语言,类似于传统的 SQL 语言。并提供 JAVA, C#, C++, PYTHON 等等常用程序语言的访问驱动。

#### (7) 数据接口:

现在可使用的接口和数据交换格式包括: OPC, Modbus, S7 TCP, API, Json, XML

### **工业应用中使用的优点**

应用此平台使得工业领域中长时间(若干年)监控数据的统一存储和分析易于实现。通过动态增删数据库节点,可对数据系统进行扩容和修复,系统可随着需求的增长而动态扩展。

面对海量数据时,工厂大数据平台可提供稳定的读写吞吐量,不会因为数据量的增加而降低性能指标。

它提供了持续可用性、高扩展性和高性能、强安全性和操作简单性,同时降低了总体拥有成本。开源的非结构化数据平台从性能和可靠性上,已经具备在工

厂环境里的应用。可以作为实时数据库的补充，为客户提供更丰富的应用环境，特别是数据分析，挖掘方面。是工厂客户搭建数据中心的可选平台。