大数据

计算机19-3刘康来2019011777

> 2018年11月，希捷公司赞助，IDC发布了最新版的白皮书《Data Age 2025》：The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big，预测了2025年全球数据量总和将高至175ZB(1 Terabyte (TB) = 1024 GB. 1 Petabyte (PB) = 1024 TB. 1 Exabyte (EB) = 1024 PB. 1 Zettabyte (ZB) = 1024 EB)。
>
> 如此大的数据正在引发IT界的重大技术变革，作为数据组织存储，访问修改，分析预测的安全易用的数据库正是这种变革的主要技术方向。

大数据?

> 来自`https://www.oracle.com/cn/big-data/what-is-big-data/`

- 大数据的定义：高速 (Velocity) 涌现的大量 (Volume) 的多样化 (Variety) 数据，3V 特性。
- 简而言之，大数据指非常庞大、复杂的数据集，特别是来自新数据源的数据集，其规模之大令传统数据处理软件束手无策，却能帮助我们解决以往非常棘手的业务难题。

新时代，新技术

传统数据库的基本架构是30年前以事务处理为主要应用设计的，不能完全满足对海量结构化和非结构化数据的存储管理、复杂分析、关联查询、实时性处理和控制建设成本等多方面的需要。因此，“多种架构支持多类应用”成为数据库行业应对大数据的基本思路。

- 数据库行业出现互为补充的三大阵营，适用于事务处理应用的OldSQL、适用于数据分析应用的NewSQL和适用于互联网应用的NoSQL。
- 以满足对于伸缩性（动态按需来）、容错性（可用性）、可扩展性（满足数据增长需求）等的需求。

大数据处理，最主要的支撑技术：

大数据的分布式和并行计算

- 分布式并行计算，将复杂任务分解成子任务、同时执行单独子任务的方法，比传统计算更快捷、更高效，在有限的时间内处理大量的数据，完成复杂度更高的计算任务。
- Hadoop，代表性的第一代开源框架，基于分布式并行计算的思想实现。
- Hadoop分布式文件系统，建立起可靠、高带宽、低成本的数据存储集群，便于跨机器的相关文件管理。
- Hadoop的MapReduce引擎，高性能的并行/分布式MapReduce算法数据的处理实现。

云计算和大数据

- 当数据的规模越来越大，存储和管理大数据，在硬件和软件上都需要提升，却成本高昂，给人民带来极大负担。云计算，提供共享计算资源集合，支持在云上进行应用程序的存储、计算、网络、开发、部署平台以及其它你能想到的流程。
- 在云计算中，所有的数据被收集到数据中心，然后分发给最终用户。而且，还有自动数据备份和恢复，给大数据管理带来极大便利。（Google云，无人能出其右）

大数据内存计算技术

- 对大数据处理能力需求，可以通过分布式计算得到基本的满足。但在想要进一步提升处理能力和速度，又需要内存计算（IMC）来完成。Hadoop之后出现了Spark，基于内存计算，大大提升了数据处理效率。
- IMC使用在主存储器（RAM）中的数据，这使得数据处理的速度更快。结构化数据存储在关系数据库中（RDB），使用SQL查询进行信息检索。非结构化数据包括广泛的文本、图像、视频等，则通过NoSQL数据库来完成存储。
- IMC处理大数据的数据量，NoSQL数据库处理大数据的多样性。

综合来看，分布式技术与并行计算依赖于硬件的提高，而数据库如何去使用，尽可能的发挥效能也是一门艺术。对云空间的数据库管理更是未来万物互联的基础，内存运算也极其重要。

> 新时代，需要新技术，新技术需要新人才。
> 数据库方兴未艾，大数据前途光明。

Reference:

```
https://blog.csdn.net/weixin_42034217/article/details/84556830
https://cloud.tencent.com/developer/news/676161
http://www.h3c.com/cn/d_201511/901094_30008_0.htm
```