什么是分布式数据库

典型回答

分布式数据库，即所谓的NewSQL。主要代表TiDB（pingcap）、OceanBase（蚂蚁）、Spanner（google）。与传统关系型数据库相比：

1. 性能：在亿级别以上，读和写都会高于传统关系型数据库（MySQL、SQLServer等），在亿级别以下略逊色与传统关系型数据库
2. 可维护性：TiDB、OceanBase都以开源的形式存在，其生态也比较贴近于现代数据库的需求（如完全支持云原生），有较好的社区文档、企业免费的培训课程
3. 可靠性：由于分布式的特性，通过副本冗余的方式提升整个集群可靠性。同样由于分布式的特性，无法严格且完全的实现真正意义上的ACID，从事务的角度来看可靠性降低了。
4. 可扩展性：分布式的特性就在于近乎无限的水平可扩展，增加集群节点数量可大幅度提高集群的QPS和存储能力。Spanner甚至实现了全球部署。
5. 用户体验：兼容大部分SQL标准，但也是由于分布式的原因，很多传统关系型数据库的特性（SQL语法、其他功能）无法支持，如：无法保证自增id的连续性，有限的支持事务的强一致性(对性能略有损失)，天生分布式不支持单机部署（小型业务无法使用）。TiDB与MySQL思想类似，也存在存储引擎概念，通过存储引起实现了OLAP和OLTP两种模式，插入数据能随时进行在线事务操作，也可以进行实时的离线分析操作。

扩展知识

有了MySQL为啥还要有分布式数据库?
其实能用上分布式数据库的公司数据量和QPS已经非常庞大了。很难通过MySQL解决数据量和高QPS的问题了。MySQL在应对大数据量的时候通过采用分库分表的方案（通常单表空间>20G，行数超过2个亿就需要进行拆分了），应对大量读QPS的场景方案是进行读写分离，而写TPS很难提高。分库分表会带来的问题：

1. 分页，聚合等操作需要中间件或应用自身支持（因此衍生出了sharding-sphere）
2. id范围做分库分表后，范围固定，数据量积累多了再扩，需要迁移数据重新分。按时间范围可能存在数据倾斜的问题。按hash分的话一样存在扩容迁移数据的麻烦事

读写分离会带来的问题：

1. 主从复制存在延迟，存在写后读依赖的场景，只能强制读主，进而对主产生了压力
2. 读写分离需要应用感知主和从的存在
3. 使用cluster方案维护主从切换成本高
4. 纯主从复制方案单实例写tps有限，从只能提高读的并发度

因此，MySQL官方提供了Cluster方案，但对主从选举时采用了复杂的paxos算法，可维护性和性能都大幅度降低，业内并不买账，依旧使用传统的分库分表、主从复制。
总的来说MySQL可维护性和可扩展性较差。

在分布式数据库中，从TiDB来说，重点解决了可维护性和可扩展性的问题。TiDB通过pd节点管理数据的自动分片，调度，倾斜自动迁移，热点自动迁移，自动选主（采用轻量的raft协议）等功能。数据节点单机使用LSM树作为底层存储结构，大幅度提升机械磁盘写的IOPS。本质上内部分片采用范围分片的策略，然后整个集群构成了一个树状结构，进行聚合，分页查询。分布式中时间是一个很重要的东西，无论是事务的id生成，还是MVCC（多版本并发控制）都依托于时间，在计算机中即使是采用NTP服务也无法解决时间漂移的问题。TiDB通过pd节点管理时间，计算节点、数据节点都从pd节点进行同步。而Spanner采用了一个硬件设备（艳原子钟）确保数据中心的时间非常精准，从而实现了全球数据库部署，即跨国际的时间同步策略保证了事务的正确。
`题外话： Leslie Lamport 在 1978年提出了分布式逻辑时钟算法，有兴趣可以读读其论文`

由于是分布式数据库，有限支持id自增，有限支持强一致的事务(对性能略有损失)。其弊端：

1. 天生分布式，不支持单机部署，基本不适合小业务量的应用
2. 由于分布式，事务隔离级别弱，无法完全实现ACID
3. 实测在亿级以下的读写请求比mysql的性能略低一些

总的来说，虽然tidb承诺金融级别的分布式数据库，但是选择用在核心链路上还要深思熟虑。单车用在支付链路上曾遇到过因为tidb 慢查询导致整个集群崩溃，其原因是tidb查询分析器有bug导致选择错了查询计划。
分布式数据库是技术和商业发展的必然，人类产生的数据量越来越大，传统关系型数据库也在逐渐往分布式数据库上过度（如MySQL的Cluster方案），人们应该逐渐接受分布式带来的问题，接受并能够解决事务不是那么强一致。

值得一读的论文：

- Google的Spanner虽然闭源，但其被奉为分布式数据库的鼻祖，TiDB大量参考了该论文的实现。有兴趣的小伙伴可以读读 《Spanner: Google’s Globally-Distributed Database》
- TiDB完全开源免费，社区非常活跃，在国内外大厂用的很多。有兴趣的小伙伴可读读《TiDB: a Raft-based HTAP database》
- Leslie Lamport 逻辑时钟 《Time, Clocks, and the Ordering of Events in a Distributed System》