x 分布式数据库

我的范文

NOSQL

特点

海量数据存储、高并发请求、高可用、高可扩展性

主要技术

列式数据库: HBase. 适合大数据量 (100TB 级数据)，而且增长量无法预估的数据; 适合写密集型应用，每天写入量巨大，而读数量相对较小的应用; 适合于批量数据处理和即时查询; 常用于联机事务型数据处理(OLTP)

键值数据库: Redis / Memcached. 适合存储用户信息（比如会话）、配置文件、参数、购物车等等与 ID 挂钩的数据。

文档型数据库: mongoDB. 文档数据库通常以 JSON 或 XML 格式存储数据. 适用于表结构不明确，且字段在不断增加.

图形数据库: neo4j. 在一些关系性强的数据应用，例如社交网络。

适用场景

1、数据模型比较简单；

2、需要灵活性更强的 IT 系统；

3、对数据库性能要求较高；

4、不需要高度的数据一致性；

5、对于给定 key，比较容易映射复杂值的环境。

背景

需求：公众出行对路况信息的需求, 交通管理部门的监控、执法、取证需求。

关系型数据库缺陷：传统关系型数据库由于由各种关系的依赖及索引的限制, 可扩展性差, 并且随着数据量的增长查询效率会急剧降低. 所以需要利用 **NOSQL 数据库**来满足海量数据的存储和索引需求。

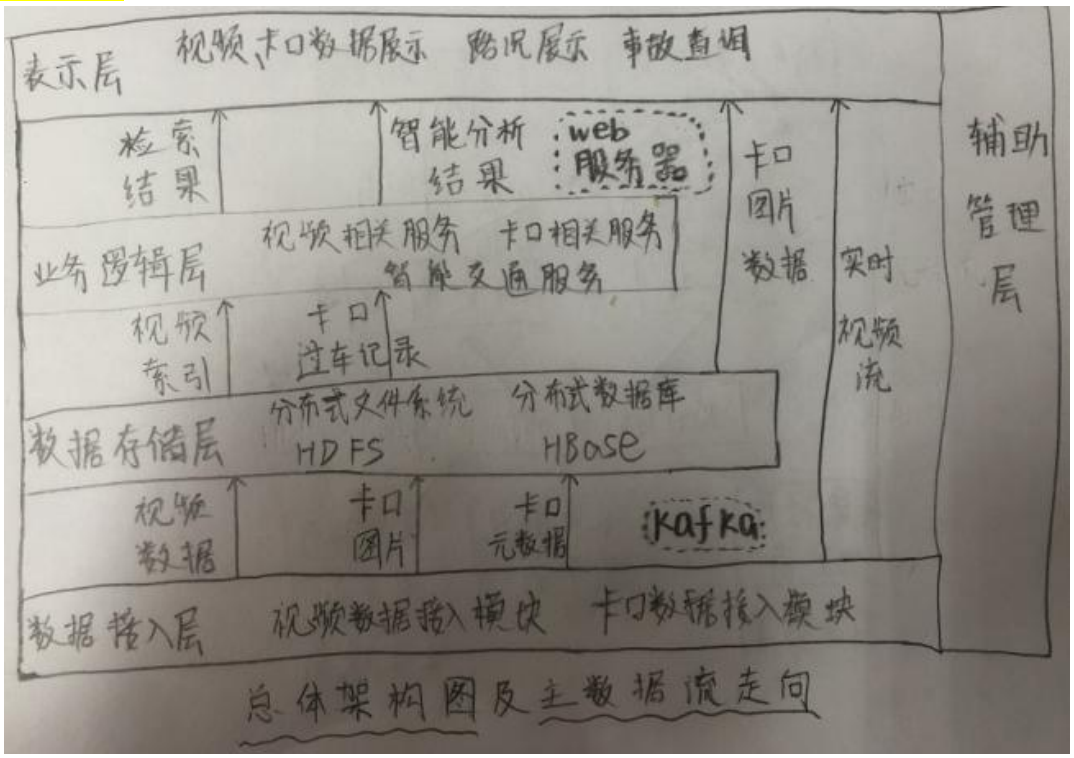
单机存储缺陷：在数据量小的时期单台机器就可以满足数据存储的需求, 随着数据量的增长, 挂载的硬盘越来越多, 发展成磁盘阵列, 但是其必须在物理上放在一起, 且只能用于存储数据, 不能用于本地计算. 不能满足大数据存储和计算的需求, 因此需要使用**分布式文件系统**解决海量数据分布式存储的需求。

需求

交通信息平台需要接入并存储100路视频监控数据以及道路监控卡口数据. 视频数据包括视频数据本身以及视频的索引, 视频的索引字段包括地点、时间, 需要支持多条件查询; 卡口数据包括卡口照片和特征数据, 特征数据包括时间、地点、车牌、车牌类型、车速、车道等信息,同样需要实现多条件查询, 且非关键字段需求支持数据的修改.

交通信息平台不但需要实现海量数据持久化存储，还需要支持用户能够按照多种维度的条件参数对数据进行检索调取。以上数据中，视频数据和卡口照片数据为非结构化数据，需要存储的数据量巨大，且需要为后期新设备的扩展提供支持，因此使用分布式文件系统 HDFS 存储；视频索引数据和卡口照片特征数据写入量巨大，数据模型比较简单，并且需要为用户提供快速的检索功能，对性能要求很高，因此用 NOSQL 技术中的列式数据库 HBase.

总体架构图



在上述需求分析的基础上，为了满足业务需求和线性扩展的要求，我建立了层次化、模块化的软件系统架构，分为数据接入层、数据存储层、业务逻辑层和表示层.

数据接入层：该层完成数据的接入，包括视频数据接入模块和卡口数据接入模块. 数据接入层可适应多种网络结构、多种形式的交通数据的接入，同时预留可扩展的接口用于未来可能出现的新接入形式.

数据存储层：该层负责完成原始数据的汇聚、存储和处理，并提供基础的检索服务. 数据存储层使用了分布式文件系统 HDFS 以及分布式数据库 HBase，实现了海量数据存储的线性扩展. 一方面对数据进行存储，另一方面对上层提供形式多样的服务.

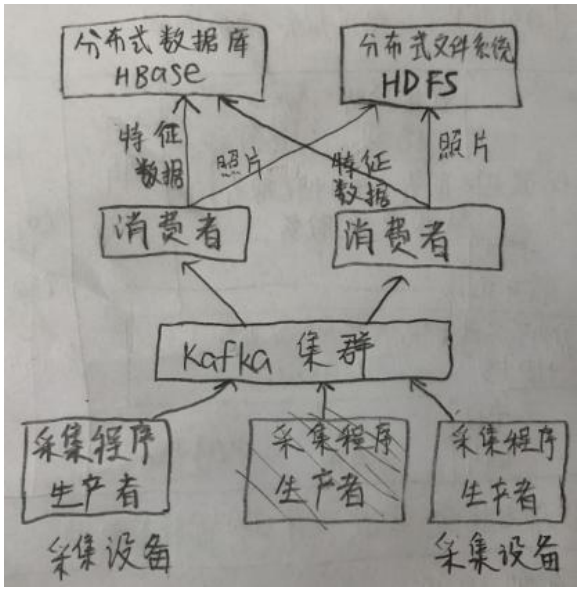
业务逻辑层：该层通过调用数据存储层的数据，通过一系列计算，为用户提供所需的服务，并交给表示层进行展示. 一方面提供视频、卡口相关数据的查询和下载服务，另一方面基于数据存储层的数据进行更深入的研判分析业务，包括路况判断、出行时间预测、交通事故预测等高级功能，同时支持第三方开发人员利用系统的接口进行二次开发。

表示层：表示层为用户提供数据展示的页面并接收用户的输入. 采用 B/S 架构. MVC...

辅助管理层: 提供用户身份认证、日志、数据加密等公共服务。

数据流如上，助于我们更好地把握系统的运行逻辑，为架构设计和接口设计提供指导.

a 卡口部分架构设计



卡口接入和存储架构图

卡口数据的特点: 1.卡口数据形式为一张照片和文本，构成了一条离散的消息; 2.卡口数据由多个客户端同时提供; 3.当数据存储平台不可用时要将这段时间的数据缓存在某个地方，不能丢失; 4.要保证数据接入的实时性，延迟控制在 1 秒以内.

数据接入设计. 针对卡口数据特点，我们使用了分布式消息队列 Kafka(分布式发布订阅消息系统)技术，建立了一个生产者-消费者模型. 首先有一个采集程序在前端设备上采集数据，并将采集的数据发送到 kafka 集群，然后有程序从 kafka 上拉取数据，进行后续持久化存储.

kafka: 独立消息工作模式/ 分布式，扩展性强，适合都客户端/ 生产者-消费者模式实现了模块间解耦/ 利用操作系统的缓存机制，读写效率高.

数据存储设计. **卡口特征数据**为包括时间、地点、车牌等信息的结构化数据，需要支持多条件查询，非关键字段需要支持数据的修改，查询后要能提取出对应的照片数据. 特征数据数据量大，但是结构清晰，不需要太多索引，数据之间也没有太强的关联关系和数据约束. 因此我们使用分布式数据库 HBase 进行特征数据的存储. **卡口照片**为非结构化数据，数据量特别大，且图片不需要做分析，只需要一次写入，多次读取，我们决定采用分布式文件系统 HDFS 进行照片的存储，这样可以将 HBase 和图片放在同一个 HDFS 集群上进行存储. HBase 中存储图片的路径，根据该路径可以在 HDFS 中取到图片.

b 视频部分架构设计

数据接入设计。视频流的数据是连续的，上一个数据包跟下一个数据包必须传给同一台机器，也就是说对于前端的一个设备来说要有一台确定的服务器跟它交互来接入视频流，此次涉及 100 多台前端视频设备，怎样合理地将前后端设备进行对应，做到服务器地负载均衡避免出现热点是我们的主要目标。为了做到负载均衡，我们设计了一个“任务协调程序”，用来协调“视频接入程序”所对应的视频设备，“任务协调程序”与“视频接入程序”绑定在一起，通过使用分布式协调系统 Zookeeper 实现集群的监控和协调工作。定期读取前端设备信息，更新到 Zookeeper 上。

数据存储设计。视频数据数据量巨大，因此数据接入后先缓存在本地内存，每一分钟存到 HDFS 上一个文件，同时将文件索引存到 HBase。这样以 1 分钟为粒度进行数据存储可以使视频的查询下载更加灵活。

历年真题

试题四

分布式存储系统（Distributed Storage System)通常将数据分散存储在多台独立的设备上。传统的网络存储系统采用集中的存储服务器存放所有数据，存储服务器成为系统性能的瓶颈，也是可靠性和安全性的焦点，不能满足大规模存储应用的需要。分布式存储系统采用可扩展的系统结构，利用多台存储服务器分担存储负荷，利用位置服务器定位存储信息，它不但提高了系统的可靠性、可用性和存取效率，还易于扩展。

请围绕“论分布式存储系统架构设计”论题，依次从以下三个方面进行论述。

1. 概要叙述你参与分析和开发的分布式存储系统项目以及你所承担的主要工作。

2. 简要说明在分布式存储系统架构设计中所使用的**分布式存储技术**及其实现机制，详细叙述你在具体项目中选用了哪种分布式存储技术，说明其原因和实施效果。

3. 冗余是提高分布式存储系统可靠性的主要方法，通常在分布式存储系统设计中可采用哪些**冗余技术**来提升系统的**可靠性**？你在具体项目中选用了哪种冗余技术？说明其原因和实施效果。

分布式存储技术

1. **集群存储技术**。集群存储系统是指架构在一个可扩充服务器集群中的文件系统，用户不需要考虑文件是存储在集群中什么位置，仅仅需要使用统一的界面就可以访问文件资源。当负载增加时，只需在服务器集群中增加新的服务器就可以提高文件系统的性能。集群存储系统能够保留传统的文件存储系统的语义，增加了集群存储系统必须的机制，可以向用户提供高可靠性、高性能、可扩充的文件存储服务。

2. **分布式文件系统**。分布式文件系统是指文件系统管理的物理存储资源不一定直接连接在本地节点上，而是通过计算机网络与节点相连。分布式文件系统的设计基于客户机/服务器模式。一个典型的网络可能包括多个供多用户访问的服务器。另外，对等特性允许一些系统扮演客户机和服务器的双重角色。分布式文件系统以透明方式链接文件服务器和共享文

件夹，然后将其映射到单个层次结构，以便可以从一个位置对其进行访问，而实际上数据却分布在不同的位置。用户不必再转至网络上的多个位置以查找所需的信息。

3. 网络存储技术。网络存储系统就是将“存储”和“网络”结合起来，通过网络连接各存储设备，实现存储设备之间、存储设备和服务器之间的数据在网络上的高性能传输。为了充分利用资源，减少投资，存储作为构成计算机系统的主要架构之一，就不再仅仅担负附加设备的角色，逐步成为独立的系统。利用网络将此独立的系统和传统的用户设备连接，使其以高速、稳定的数据存储单元存在。用户可以方便地使用浏览器等客户端进行访问和管理。

4. P2P 网络存储技术。P2P 网络存储技术的应用使得内容不是存在几个主要的服务器上，而是存在所有用户的个人电脑上。这就为网络存储提供了可能性，可以将网络中的剩余存储空间利用起来，实现网络存储。人们对存储容量的需求是无止境的，提高存储能力的方法有更换能力更强的存储器，另外就是把多个存储器用某种方式连接在一起，实现网络并行存储。相对于现有的网络存储系统而言，应用 P2P 技术将会有更大的优势。P2P 技术的主体就是网络中 Peer，也就是各个客户机，数量是很大的，这些客户机的空闲存储空间是很多的，把这些空间利用起来实现网络存储。

分布式存储系统设计中的冗余技术

1 数据备份

2 数据分割

3 门限方案

4 纠错编码和纠删编码。纠删编码---把丢失的数据计算出来；

以**数据备份**为例。Hadoop 视硬件错误为常态，并通过块的冗余存储机制保证数据的高可靠性。在大多数情况下，副本系数是 3，HDFS 的存放策略是将一个副本存放在本地机架的节点上，一个副本放在同一机架的另一个节点上，最后一个副本放在不同机架的节点上。当某一个节点的硬件失效或者数据块异常时，则访问其它节点正常的块数据副本~~，并尝试将正常块复制到失效节点上，以恢复异常块。~~

试题三

大规模分布式系统通常需要利用缓存技术减轻服务器负载、降低网络拥塞、增强系统可扩展性。缓存技术的基本思想是将客户最近经常访问的内容在缓存服务器中存放一个副本，当该内容下次被访问时，不必建立新的数据请求，而是直接由缓存提供。良好的缓存设计是

一个大规模分布式系统能够正常、高效运行的必要前提。在进行大规模分布式系统开发时，必须从一开始就针对应用需求和场景对系统的缓存机制进行全面考虑，设计一个可伸缩的系统缓存架构。

请围绕“大规模分布式系统缓存设计策略”论题，依次从以下三个方面进行论述。

1. 概要叙述你参与实施的大规模分布式系统开发项目以及你所担任的主要工作。

2. 从不同的用途和应用场景考虑，请详细阐述至少两种常见的**缓存工作模式**，并说明每种工作模式的适应场景。

3. 阐述你在设计大规模分布式系统的缓存机制时遇到了哪些问题，如何解决。

缓存工作模式

单实例缓存模式(Single Instance)、

复制模式(Replication Cache)

分区模式(Partition Cache)。

(1) 单实例模式。单实例模式是一种较为简单的缓存模式，多个应用服务器共享一个中央的缓存服务器。通过共享缓存的数据，能够极大地提高系统的性能。该模式的主要限制在于缓存服务器的内存大小和节点增加之后服务器的处理能力和网络带宽。该模式的适用场景是：对缓存的要求比较简单；系统的吞吐量和数据量不大；性能要求不高。

(2) 复制模式。复制模式将缓存的数据复制到多台机器上（多实例），对于单一缓存服务器性能出现问题的情况下，可以通过缓存复制的方式将压力分解到多个缓存服务器。该模式的工作原理是：缓存客户端可以访问自己的缓存服务器，多个缓存服务器之间的数据是彼此同步的，对于性能要求更高的场景，这样的部署架构能够获得更高的吞吐能力。该模式的适用场景是：数据量不是特别大；需要极高的性能；数据改动的频率不是特别大。

(3) 分区模式。当需要缓存的数据已经超过一台服务器的内存上限时，可以考虑采用分区模式对数据进行线性缩放，也就是通过增加缓存服务器来解决数据增长和压力增加的情况。在分区模式中，其架构是无分享架构（SharedNothing Architecture，SNA)，每个节点之间数据彼此独立，一个节点出现故障后不会影响到其他节点。在出现某个节点宕机或者其他故障的情况下，致使这部分的分区缓存无法使用，并不妨碍其他数据节点数据的正常工作。该模式的适用场景是：总体数据量较大。已经超出了单个缓存服务器的内存上限；系统缓存要求具有很大的可伸缩性；客户端量庞大，单个客户端对缓存数据的数据量要求不大。

缓存机制设计时遇到的问题

如何缓存服务器的工作模式选择；

高可用性的设计考虑；

缓存一致性与分布式算法；

对象状态同步的考虑；

缓存钝化、激活、过期和初始化　等