OSC源创会 Opensource Innovation Meetup | IT大咖说 不止于技术

分布式数据库灵活存储机制的实现与应用

乔国治 技术总监
qiaoguozhi@sequoiadb.com

SequoiaDB 巨杉数据库

公司简介：

SequoiaDB巨杉数据库（广州巨杉软件），成立于2011年专注于新一代企业大数据平台研发, 其核心产品SequoiaDB（巨杉数据库）是国内第一款新一代分布式数据库;

核心产品完全自主研发，数据库引擎没有基于任何开源数据库源代码, 已经成功部署并运行在多家世界500强企业的生产环境中;

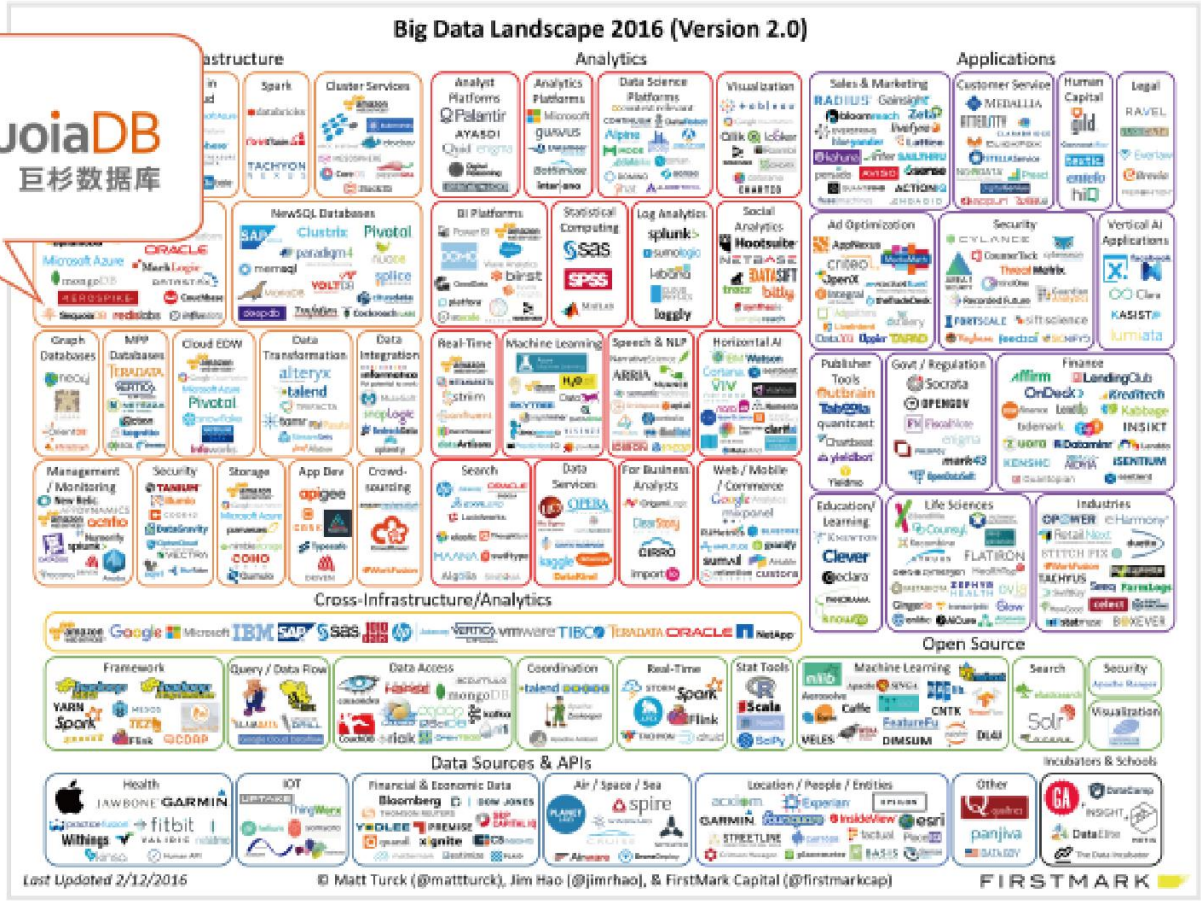
获著名基金启明创投（A轮）与DCM（B轮）融资;

中国第一款商业开源数据库产品
www.github.com/sequoiadb/sequoiadb
www.sequoiadb.com

"全球创新企业Top100"
——《红鲱鱼》
美国最具影响力商业媒体

"中国创新企业50强"
——《快公司》
美国著名创新媒体

SequoiaDB
巨杉数据库



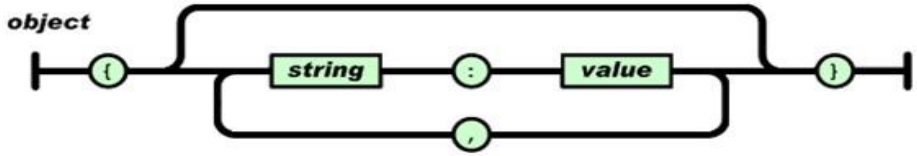
成熟的商业化自主研发数据库 — 行业用户认可

- 主要客户以金融、运营商、政府、交通航空、互联网等行业为基础
- 研发中心在深圳，现场支持队伍部署在北京、上海、广州三地

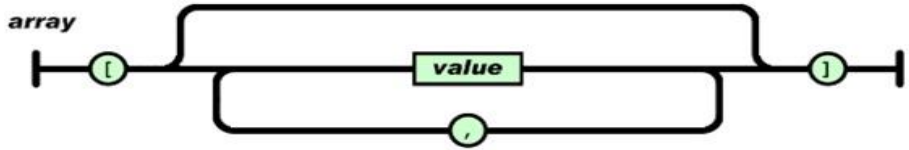
JSON/BSON 记录存储

JSON（JavaScript Object Notation）是一种轻量级的数据交换格式，它基于ECMAScript的一个子集，为纯文本格式，支持嵌套结构与数组。

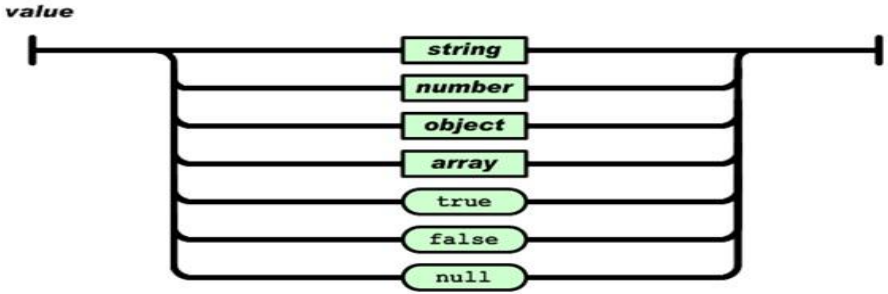
1、对象是一个无序的“键值对”集合，以“{”（左大括号）开始，“}”（右大括号）结束。每一个元素名后跟一个“:”（冒号）；而元素之间使用“,”（逗号）分隔；



2、数组是值的有序集合，以“[”（左中括号）开始，“]”（右中括号）结束。值之间使用“,”（逗号）分隔；

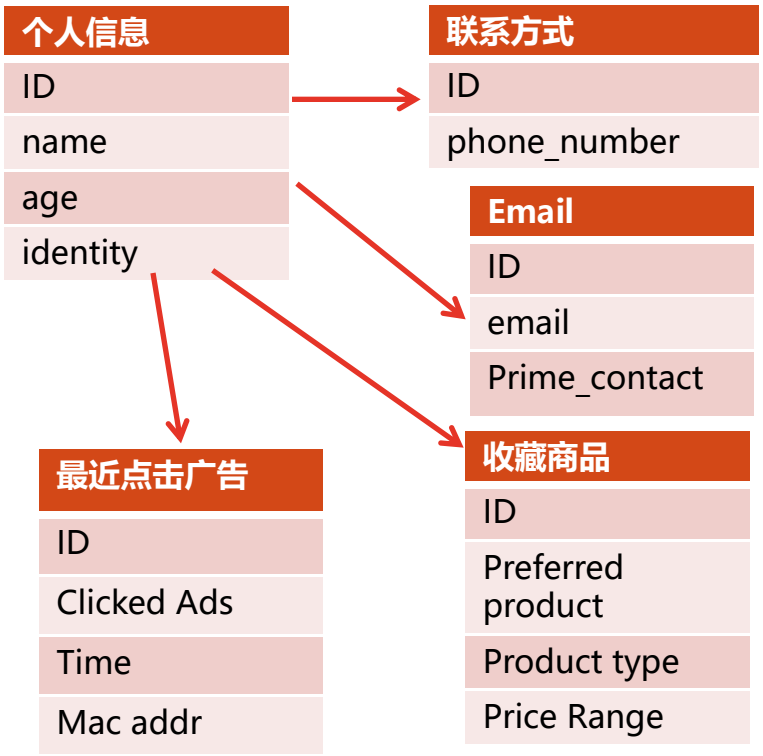


3、值可以为由双引号包裹的字符串，数值，对象，数组，true，false，null，以及 SequoiaDB 数据库特有的数据结构（例如日期，时间等）组成。



去范式的场景：对象存储

层次化嵌套形式实现一对多模型转换



个人信息	
ID	
name	
age	
identity	
电话列表[…]	phone_number
Email列表[…]	Email [email1, email2, …]
	Prime_contact
收藏商品列表 […]	Preferred product
	Product type […]
	Price Range
最近点击广告列表[…]	Clicked Ads
	Time
	Mac Addr

简化关系模型中的关联关系

Customer ID	First Name	Last Name	City
0	John	Doe	New York
1	Mark	Smith	San Francisco
2	Jay	Black	Newark
3	Meagan	White	London
4	Edward	Daniels	Boston

Account Number	Branch ID	Account Type	Customer ID
10	100	Checking	0
11	101	Savings	0
12	101	IRA	0
13	200	Checking	1
14	200	Savings	1
15	201	IRA	2

```
{     customer_id : 1,
      first_name : "Mark",
      last_name : "Smith",
      city : "San Francisco",
      accounts : [    {
            account_number : 13,
            branch_ID : 200,
            account_type :
"Checking"
      },
      {     account_number : 14,
            branch_ID : 200,

      account_type : ”Savings”,
            beneficiaries: [...]
      } ]
}
```

关系型数据结构

文档型数据结构

历史数据中同一个表的不同年份数据需要放在一起

例子：社保+商业保险的关联汇总

每年的数据表结构都会变化，历史贴源数据要包容变化的数据结构。

时间	IDh	年收入（元）	累计标保和	缴付保费合计	关联帐户	寿险缴付金额	两全缴付金额	年金缴付金额	医疗缴付金额	意外缴付金额	交通缴付金额

时间	IDh	年收入（元）	累计标保和	缴付保费合计	关联帐户	寿险缴付金额	两全缴付金额	年金缴付金额	医疗缴付金额	意外缴付金额

时间	IDh	年收入（元）	累计标保和	缴付保费合计	关联帐户	寿险缴付金额	两全缴付金额	年金缴付金额

需要每年动态增加

时间	IDh	年收入（元）	累计标保和	缴付保费合计	关联帐户	寿险缴付金额	两全缴付金额	年金缴付金额	医疗缴付金额	意外缴付金额	交通缴付金额
2012	XX47	48200	26900	80400	A1212	500	5000	0	Null	Null	Null
2012	XX47	48200	26900	80400	A1213	0	0	76000	Null	Null	Null
2012	XX51	29400	15000	96600	A2039	500	90400	5800	Null	Null	Null
2013	XX40	180600	10800	21300	A3359	1100	3800	12400	100	800	Null
2013	XX37	29400	7400	71100	A6596	600	5100	1500	100	500	Null
2014	XX33	28200	6500	30700	A8767	200	800	29300	100	100	300
2014	XX53	33600	5800	80500	A9785	200	5300	2500	100	200	200
2015	XX42	97400	5300	9800	A9078	400	3300	3000	100	300	400

历史数据的动态存储方式

时间	IDh	年收入（元）	累计标保和	缴付保费合计	关联帐户	寿险缴付金额	两全缴付金额	年金缴付金额
2012	XX47	48200	26900	80400	A1212	500	5000	0
2012	XX47	48200	26900	80400	A1213	0	0	76000

```
{     Time : 2012,
      IDh : “xx47”,
      Income : 48,200,
      Acc_Sum : 26,900,
      Invoice_Sum : 80,400,
      accounts : [    {
            account_number : A1212,
            life_insurance: 500,
            labor_insurance: 5000,
            pension:0,
      },
      {     account_number : A1213,
            life_insurance: 0,
            labor_insurance: 0,
            pension:76,000,
      } ]
}
```

时间	IDh	年收入（元）	累计标保和	缴付保费合计	关联帐户	寿险缴付金额	两全缴付金额	年金缴付金额	医疗缴付金额	意外缴付金额	交通缴付金额
2014	XX33	28200	6500	30700	A8767	200	800	29300	100	100	300

```
{     Time : 2014,
      IDh : “xx33”,
      Income : 28,200,
      Acc_Sum : 6,500,
      Invoice_Sum : 30,700,
      accounts : [    {
            account_number :
A8767,
            life_insurance: 200,
            labor_insurance: 800,
            pension:29300,
            health_care:100,
            accident: 100
            transportation:300
      }]
}
```

文件存储 与 LOB机制

BSON和LOB按需搭配使用

缩略图 - BSON Binary

•{id:”001”,time:”2015120
1”,...photo:”aGVsbG8gd29ybGQh”}

•Binary只是BSON中的一个字段类型，以行的方式直接存储。对于数据库来讲，这就是一条普通记录。所以访问时不会产生额外的IO。

高清图 - LOB

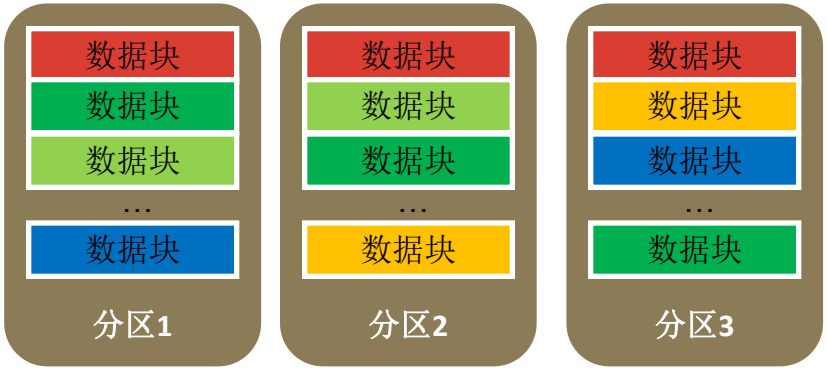
•单独的分布式块存储模式

•文件按照LOB处理

•自动按照64/128KB的数据块进行切分，放在不同分区存储

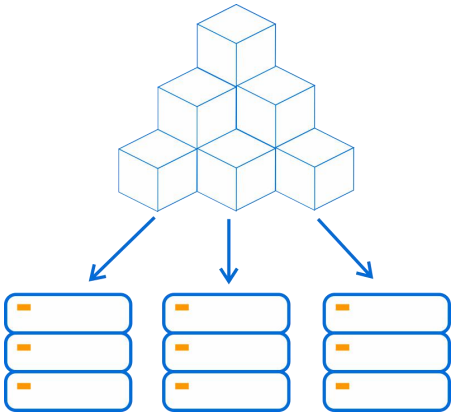
•使用DIO避免二进制数据占用文件系统缓存

•并行处理



LOB 块存储机制

与其他解决方案相比，由于不存在独立中控元数据节点，SequoiaDB提供的LOB存储机制理论上可以存放近乎无限数量的对象文件，并且不会由于元数据堆积而造成性能下降。同时，由于数据块被散列分布到所有数据节点，整个系统的吞吐量随集群磁盘数量的增加近乎线性提升。



文件按照LOB处理

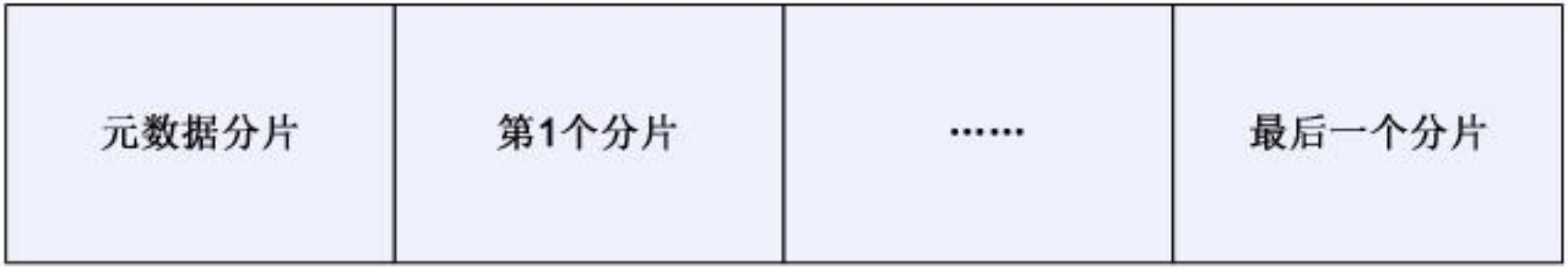
- 自动按照64/128KB的数据块进行切分，放在不同分区存储
- 使用DIO避免二进制数据占用文件系统缓存
- 并行处理

- 与GridFS相比不占用内存
- 与HDFS相比不存在Namenode限制

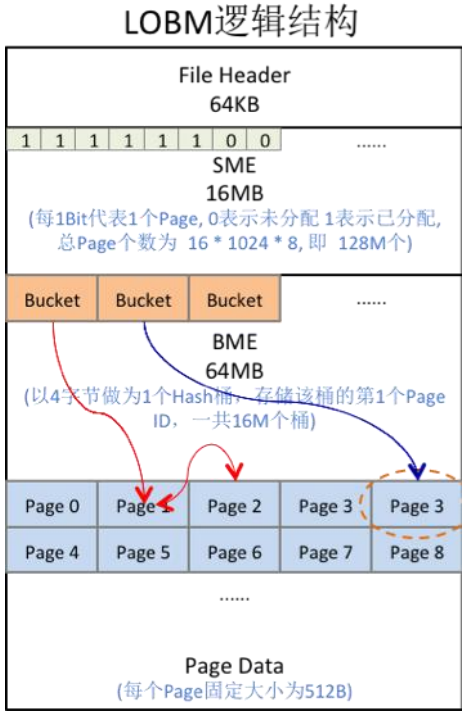
LOB 块存储机制

大对象（LOB）功能旨在突破 SequoiaDB 的单条记录最大长度为 16MB 的限制，为用户写入和读取更大型记录提供便利。LOB 记录的大小目前不受限制。

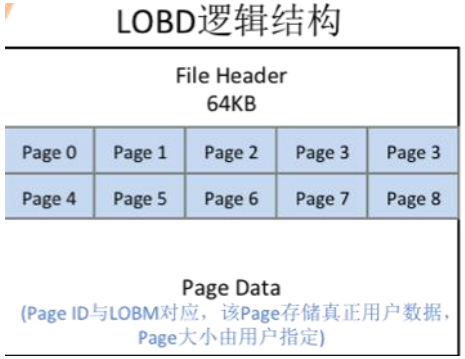
每一个 LOB 记录拥有一个 OID，通过指定集合及 OID 可以访问一条 LOB 记录。在非分区集合及哈希分区集合中均可使用 LOB 功能。集合间不共享 LOB 记录。当一个集合被删除时，其拥有的 LOB 记录自动删除。
LOB 记录的存储格式：



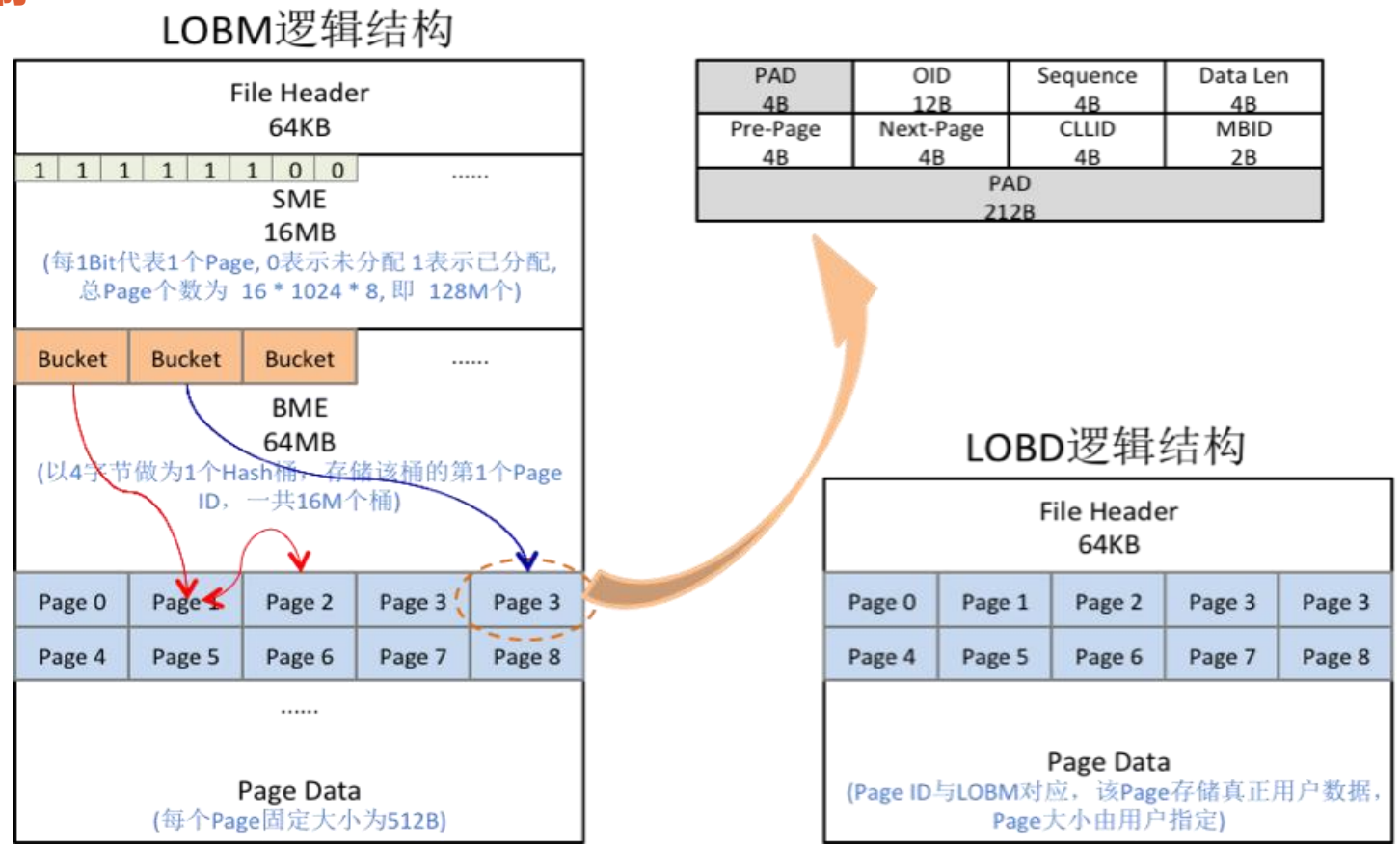
LOB 块存储机制



SequoiaDB的LOB存储结构分为元数据文件（LOBM）与数据文件（LOBD）。
其中，元数据文件存储整个LOB数据文件的元数据模型，包括每个页的空闲状况、散列桶、以及数据映射表等一系列数据结构。而数据文件则存储用户真实数据，数据头之后所有数据页按照page size进行切分，每个数据页不包含任何元数据信息。在建立集合的过程当中，大对象存储必须依附于普通集合存在，一个集合中的大对象仅归属于该集合，不能被另外一个集合管理。



LOB 映射机制


LOBM逻辑结构
File Header
64KB
1 1 1 1 1 1 0 0
.....
SME
16MB
(每1Bit代表1个Page, 0表示未分配 1表示已分配,
总Page个数为 16 * 1024 * 8, 即 128M个)
Bucket
Bucket
Bucket
.....
BME
64MB
(以4字节做为1个Hash桶，存储该桶的第1个Page
ID，一共16M个桶)
Page 0
Page 1
Page 2
Page 3
Page 3
Page 4
Page 5
Page 6
Page 7
Page 8
.....
Page Data
(每个Page固定大小为512B)
PAD
4B
OID
12B
Sequence
4B
Data Len
4B
Pre-Page
4B
Next-Page
4B
CLLID
4B
MBID
2B
PAD
212B
LOBD逻辑结构
File Header
64KB
Page 0
Page 1
Page 2
Page 3
Page 3
Page 4
Page 5
Page 6
Page 7
Page 8
Page Data
(Page ID与LOBM对应，该Page存储真正用户数据，
Page大小由用户指定)

记录/文件 双存储引擎的应用实践

巨杉对象存储机制：兼顾海量小文件和大文件的存储特性

- 其他实现方式
- 关系数据库+文件系统地址

问题：文件条目受关系型数据库的性能限制，比如超过3亿条后性能急剧下降

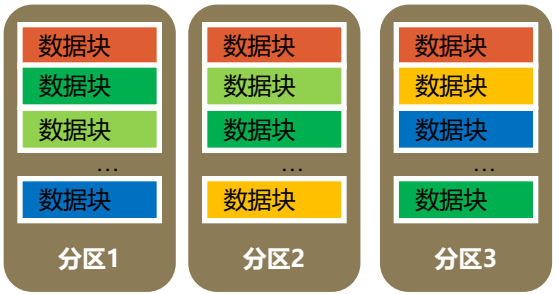
- HDFS

问题：受 Namenode限制无法处理大量的小文件，分配64MB存储浪费空间；小文件不定期后台自动聚合，影响系统的使用稳定性。

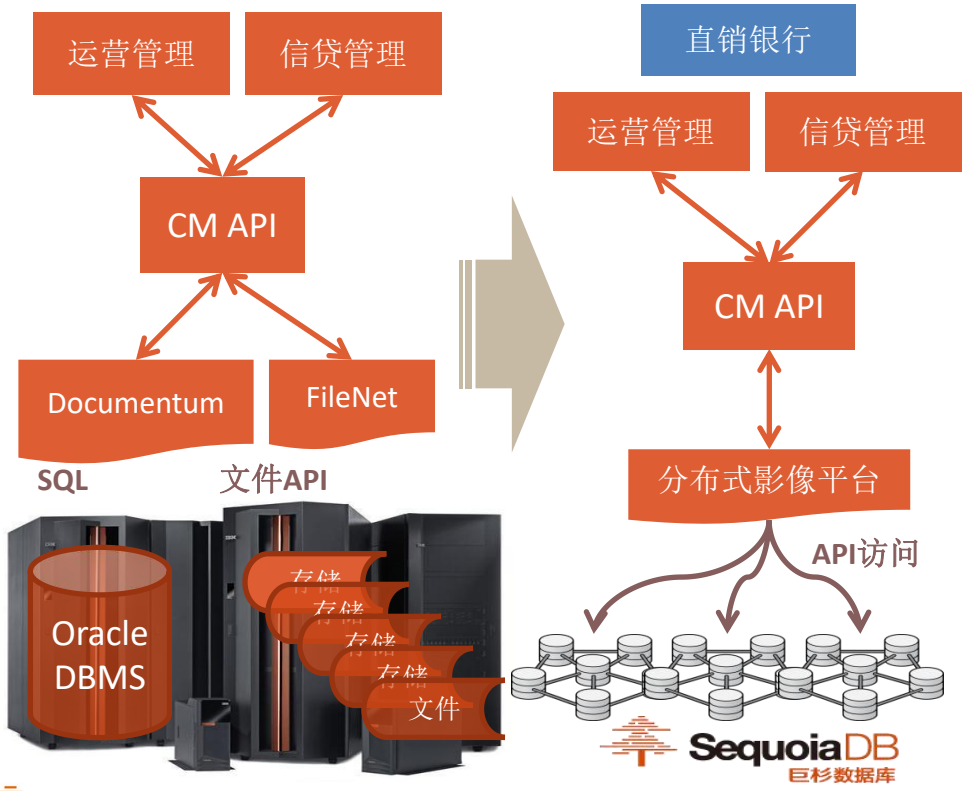
- HBase

问题：做Merge的过程造成I/O飙高，无法满足在线ECM服务场景

- 分布式对象存储
- 文件按照数据块处理
  - -自动按照64/128KB的数据块进行
  - -切分，放在不同分区存储
- 使用DIO避免二进制数据占用文件系统缓存并行处理
  - -与GridFS相比不占用内存
  - -与HDFS相比不存在Namenode限制



金融行业 - 影像平台



问题和目标

- 随着数据量增长FileNet和Documentum性能急速下降（超过2亿记录）
- 以X86低成本存储的扩展来应对不断加速增长的数据存储和实时查询需求
- 降低数据存储成本和Documentum, FileNet和Oracle的许可证成本。

需求

- 提高扩展性和稳定性，降低存储成本
- 10亿记录的并行实时查询性能
- 支持~250TB数据量，单个文件平均100K, 最大10M, 少量数据>16M
- 灵活增加影像的新描述项，提供多索引查询检索

实现方式

- 先替换文件存储，再把Oracle数据库做迁移
- SDB对外提供SQL接口，可继续使用现有查询应用

地方政府 - 政务数据大数据湖

- 一卡通项目中电子证照数据的共享

医疗行业---大数据知识库/影像数据

SequoiaDB的分布式架构支持了海量的数据的存储，同时其JSON/LOB的架构又可以满足非结构化数据的存储。这两个特性可以说是医疗行业对于数据的核心需求。

主要需求：

•海量非结构化半结构化数据：处方，药方，病历，X光片等

•数据结构多样化：每个医生、患者对应的数据结构都不同，难以统一

•数据量大，历史数据需要实时在线查询：诊断历史数据的实时查询，病情跟踪

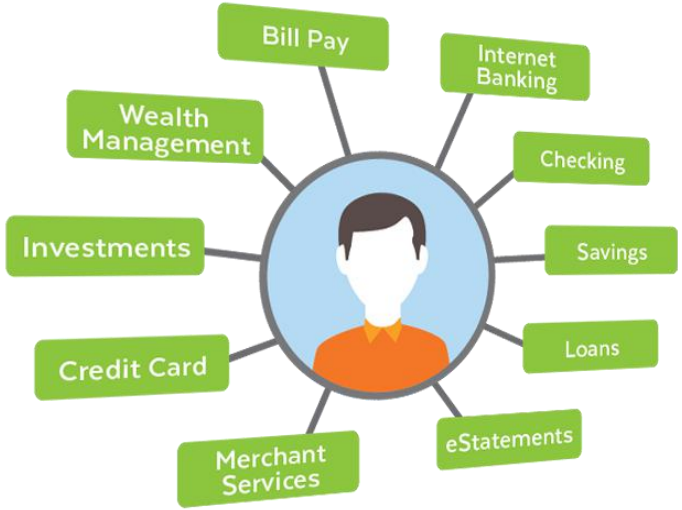
OTA旅游 /电商 - 多类型数据混合存储

互联网应用的特点，带来了几项重要的挑战。
•数据量大
•业务增长快
•数据类型多样

途牛旅游网“资源系统”的另一个核心业务模块。其负责存储和记录所有的旅游方案相关的资源信息，包括酒店，机票，门票，火车票，汽车票，地接，当地服务等，和价格中心相同，我们的资源也呈海量存储的特点，同样在对于静态资源的读取上，通过和巨杉的合作也

360度用户视图

通过大数据对用户行为进行分析，已经成为各行各业对大数据应用场景的基本认知。不论是金融、政府、运营商、制造、甚至互联网等行业，都在考虑如何使用大数据技术，借助用户行为分析、第三方数据分析等方式，进一步完善已有的CRM体系，将传统的静态数据向360度用户视图转移。



免税易购应用场景简单介绍

免税易购 享免税 乐生活 mianshui365.com
HNDF 海南免税(海口美兰机场免税店)唯一官方网上预订平台

支撑

海南离岗免税美兰机场店

寻找目标客户群

- 针对美兰机场离岗的乘客进行潜在客户营销
- 针对全岛进港的人群进行潜在客户营销

寻找合适的推广投放方向

- 商品热度调整商品排列顺序
- 互联网广告投放效益判断及调整

寻找优质客户

- 频繁往返海南的有消费能力的客户
- 往返海南并形成交易习惯的客户
- 往返海南的微博名人&论坛名人

客户行为分析

- 电商平台客户访问习惯&热点商品
- 电商交易热点商品客户特性
- 电商&门店客户交易品类差异

交通/安防 监控视频影像管理

在公安与交通行业，针对视频卡口的大数据存储、分析与应用一直以来是最受关注的主题。借助SequoiaDB半结构化对象存储、分布式横向扩展能力以及非结构化影像存储引擎，交通部门可以从卡口视频文件中提取出的车牌信息、位置信息、以及时间信息按照三个维度汇总，进行道路拥堵预测、车辆轨迹跟踪、套牌车监控、尾随车辆监控等多种安防措施。

绑架事件跟踪

◎ 以被绑架人的车辆信息为切入点，查询事件发生前两周内，哪些卡口发现被绑架人的车辆；

◎ 查询这些卡口在被绑架人车辆通过后的2分钟内所有车牌信息，形成数据集；

◎ 数据集的交集信息就可能是嫌疑人的跟踪车辆。

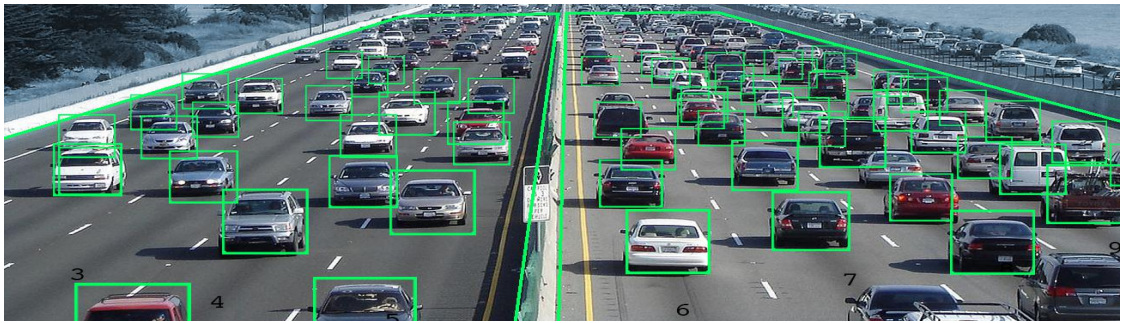
嫌疑车辆排查

◎ 出没地点排查，缩小搜索范围；

◎ 查询在多日中，嫌疑车辆深夜最后出现的卡口地点以及第二天首次出现的卡口地点。多个卡口的地点信息形成一个缩小的范围。

套牌车检测

◎ 间隔较短的时间内，在相距较远的卡口出现同一车辆号牌，有可能是套牌车产生预警，记入关注信息数据库。



SequoiaDB 2.8 即将发布
关注微信号获取最新推送

SequoiaDB 巨杉数据库
www.sequoiadb.com
Sales_support@sequoiadb.com
hr@sequoiadb.com
400-8038-339

招聘技术大牛

北京/上海/广州/深圳
数据库研发/大数据开发/售前技术支持