分布式文件存储系统的设计

**文档历史**

|  |  |  |
| --- | --- | --- |
| 版本 | 日期 | 修改内容 |
| V1 | 2013-07-30 | 初始文档，编写数据模型的实现部分 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

目录

[第一章 数据模型 1](#_Toc362959035)

[1.1 实体 1](#_Toc362959036)

[1.1.1 Column 1](#_Toc362959037)

[1.1.2 ColumnFamily 2](#_Toc362959038)

[1.1.3 KeySpace 2](#_Toc362959039)

数据模型

## 实体

主流的NoSQL存储系统实现有两大类：

• Key-Value Based Storage (源自于 Amazon's Dynamo)

• ColumnFamily Based Storage (源自于 Google's BigTable)

Key-Value Based 的系统简单到可以描述为：

|  |
| --- |
| v = storage.get(k);  storage.put(k, v);  storage.delete(k); |

因为 Key-Value 本身是作为通用数据存储系统而诞生的，所以不花篇幅讨论，这里讨论的是 ColumnFamily Based 系统，支持 Column/ColumnFamily 是为了应对复杂的业务需求：复杂的查询与特定的排序。

### Column

列 (Column) 为 ColumnFamily Based 的系统中最小的原子单元（atomic unit ), 由name, value 和timestamp组成

|  |  |
| --- | --- |
| Column | |
| Field | Type |
| Name | byte[] |
| Value | byte[] |
| Version | long(timestamp) |

**实例：**

一个用来存储 User 的 username 的列

|  |  |  |
| --- | --- | --- |
| username | | |
| Name | Value | Timestamp |
| “username” | “alpha” | 1372572814000 |

\* 1372572814000为 2013年6月30日 下午2:13:34毫秒数的UNIX时间戳。

### ColumnFamily

大体上，ColumnFamily是为了将相关的列群组在一起，以便一次性读出相关数据，而不是单个分散的数据。

Key 决定从哪里找，怎么找到对应的ColumnFamily（见排序）。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ColumnFamily | | | | | | | |
| key | Column List (1…\* Columns) | | | | | | |
|  | Name | Value | Timestamp | Name | Value | Timestamp | ... |

**实例：**

一个ColumnFamily用于存放用户的信息：username, age, gender, …（因为空间有限，只列出前两列）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| User | | | | | | | |
| key | username | | | age | | | ... |
| Name | Value | Timestamp | Name | Value | Timestamp |  |
| 1 | “username” | “alpha” | 1322105600386 | “age” | 18 | 1356105600381 |  |
| 2 | “username” | “beta” | 1332105600386 | “age” | 22 | 1356105600383 |  |
| 3 | “username” | “gamma” | 1352105600386 | “age” | 25 | 1356105600370 |  |
|  |  |  |  |  |  |  |  |

\* 列的最终组织方式是根据 ColumnFamily 指定的排序类型确定（见排序）

### KeySpace

KeySpace 是顶级的数据单元，可以想象成 RDBMS的 Schema，一个分布式存储中可以有一个或同时有多个 KeySpace。因此，同一个存储系统就可以被多个不同的应用共享。

KeySpace 包含 1..\* 个 ColumnFamily。