第2节我们来看一下子HBase数据模型，这节我们讲的内容有数据模型的相关概念，数据的坐标，概念视图，物理视图以及面向列的存储。

我们知道HBase是一个稀疏多维度排序的映射表，它的索引只有行键、列族、列限定符和时间戳，并且它的数据类型是没有的，只用字符串来表示每个值，用户可以自行进行数据类型的转换。Hbase当中存储的数据，每一行都有一个可排序的行键和任意多的列，所以表在水平方向是有一个或者多个列族组成的，那么同一个列族里面的数据是存储在一起，所以Hbase具有动态扩展强，不需要预先定义列的数量以及类型，而在更新操作的时候呢，也不会删除旧版本当中的这个旧值，而是会生成一个新的版本，旧的版本的值仍然保留。

接下来我们看一下子HBase当中数据模型的几个相关概念，首先看一下什么是表？它是Hbase采用表来组织数据，由行和列组成，列划分为若干个列族。第二个概念是行，它是每个Hbase表中若干行的组成，由行键来标识的。第3个概念呢，是列族，它是一个Hbase表为分组成许多列族的集合，它是基本访问控制单元。如右图中的Info。第4个概念是列限定符，列族里的数据，通过列限定符来定位的，比如右图当中的name， major， EMAIL。第5个概念是单元格，在Hbase当中，通过行、列族和列限定符确定的一个单元格，而单元格中存储的数据是没有数据类型的，他总是认为是字节数组byte[],如右图所示。最后一个概念是时间戳，是每一个单元格都保存着同一份数据的多个版本，这些版本是靠时间戳进行索引的，也是通过时间戳来区分每一个版本的数据，如右图当中的ts1，ts2就是两个版本的时间戳。

接下来我们看一下的数据坐标，在HBase当中，我们知道要定位一个单元格是靠行键、列族、列限定符和时间戳来确定的，所以我们就可以来把这4个值作为一个定位坐标，称为“四维坐标”，也就是[行键, 列族, 列限定符, 时间戳]，通过这四个值我们就能取得键相对应的值，比如在下图中对应的两个键值情况，我们把它列在一个键和值相对应表格当中。如[“201505003”, “Info”, “email”, 1174184619081]键的值为“xie@qq.com”。

我们再来看一下子概念视图，HBase当中的概念视图如下表所示，只给出了一个行键，在这个行键“com.cnn.www”中，有两个列族，第1个列族是contents包含有一个列限定符html，另外一个列族是anchor，包含有两个列限定符，一个是cnnsi.com，另一个是my.look.ca。这两个列族中，对应的一个时间戳在t4，t5的时候，列族contents中是没有值，而在t1，t2，t3这个时间戳，列族anchor里面是没有值的。从这个行键上，我们能看到，这里存储的是网页信息，同时也能够看到这个表当中有几个时间戳是没有值的，显然它是一个稀疏的表。

我们从概念视图当中可以看到HBase的存储是一个稀疏的表，但实际物理存储时不是这样的稀疏存储，那如何存储呢？就是说把行键先从HBase数据中单独拿出来，然后去存储好，我们再把行进行时间戳和列族中给他们抽出来，就构成了我们上面这个表，这个就是底层在物理存储中存在底层的实际的这个表，所以你看出来我们再把另外一个时间戳和列族组合起来，就构成了我们下面这个表底层存储的情况，实际上是存储这么两个时间戳，它并没有存储在一起，对于一些空格，这个当中就不存了，这些都是有数据的存储，所以我们说呢，从概念到如何物理存储思维角度来讲，它实际上是有区分的，大家通过这么两个表，最后看出它的区分。

接下来我们重点介绍一下面向列的存储，我们知道对传统的关系数据库它的存储都是以行优先的方式来进行存储的，而我们现在的列式数据库，是以列优先的顺序来进行存储的。

行式存储有它的优势，也就是一行存完之后再存第2行，那么在这一行当中就可以包含很多个列，比如说表中的Log_Id、user、age、sex列等等，这些数据都在一起，当要查询某一行数据时，行式存储是很有利的。

行式存储是一行一行的存储的，所以查询后我们能够得到图中的多个行，他们前三行的姓名列分别为mary 、Bob和tom，这样就方便查询某一行数据的所有列数据都可以查询出来。

如果我们把它前面这个表转换成列的方式来存储的话，我们如图所示，那么第1行全部存储是user一列的值，2行、3行、4行、5行分别对应着列age、sex 、Ip、 action，所有列值都存储在一起，这样有什么好处呢？大家可以思考一下。显然对于分析某一列数据时是非常有利的。在实际应用中，数据分析只对ip列有兴趣，只想得到这些数据，如果按列式存储，查询数据的速度是

非常快的，有利于大数据的处理。