# Azure DocumentDB 中的自动索引

本文摘自[“Azure DocumentDB 不限架构的索引”](http://www.vldb.org/pvldb/vol8/p1668-shukla.pdf)一文，这篇文章于 2015 年 8 月 31 日至 9 月 4 日在[超大型数据库第 41 届内部会议](http://www.vldb.org/2015/)上发布，介绍了 Azure DocumentDB 索引工作原理。

在阅读本文之后，你将能够回答以下问题︰

* DocumentDB 如何从 JSON 文档中推断架构？
* DocumentDB 如何跨不同文档建立索引？
* DocumentDB 如何执行大规模自动索引？

## DocumentDB 索引的工作原理

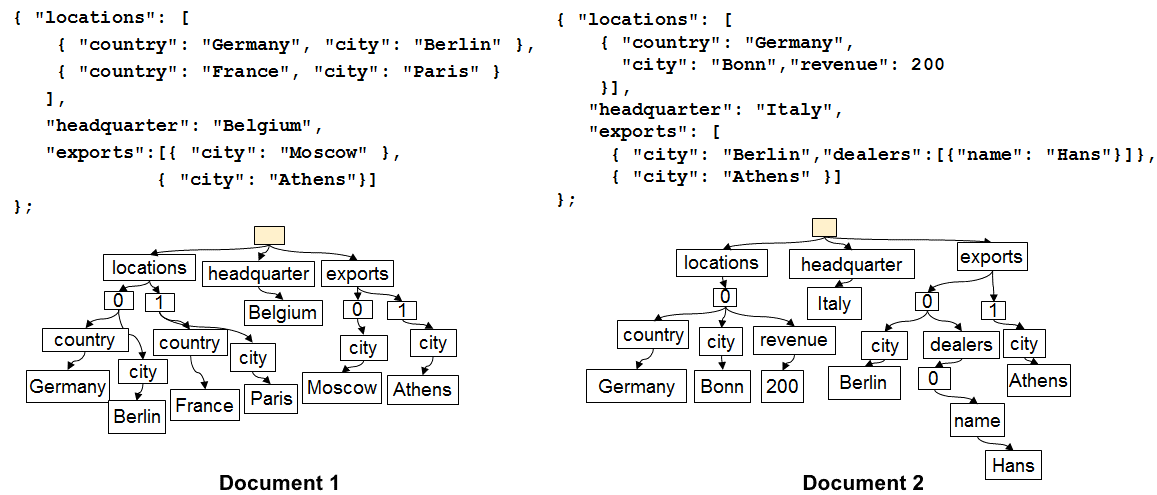
[Microsoft Azure DocumentDB](/services/documentdb/) 是专为 JSON 构建的真正的架构自由的数据库。它不需要任何架构或辅助索引定义来实现大规模的数据索引。这样，就可以使用 DocumentDB 快速定义并循环访问应用程序数据模型。将文档添加到集合时，DocumentDB 会自动为所有文档属性编制索引，以供查询时使用。借助自动索引，你可以存储属于完全任意架构的文档，而无需担心架构或辅助索引。

为了消除数据库和应用程序编程模型之间的阻抗失配，DocumentDB 利用 JSON 的简洁性以及没有架构规范的优势。它不对文档作任何假设，除了实例的具体值以外，还允许 DocumentDB 集合中的文档改变架构。与其他文档数据库相比，DocumentDB 的数据库引擎直接在 JSON 语法级别工作，仍然不受文档架构概念的限制，并且打破了文档的结构和实例值之间的界限。进而使它能够自动索引文档，而无需架构或辅助索引。

DocumentDB 中的索引利用 JSON 语法允许文档**以树形表示**这一事实。对于以树形表示的 JSON 文档，需要创建虚拟根节点，该节点是文档中下面的其余实际节点的父节点。JSON 文档中包含数组索引的每个标签都将成为树中的节点。下图说明了一个示例 JSON 文档及其对应的树形表示形式。

[AZURE.NOTE] 由于 JSON 可以自我描述，即每个文档中都包含架构（元数据）和数据，例如 {"locationId": 5, "city": "Moscow"} 表示有两个属性 locationId 和 city，并且这两个属性具有数值和字符串类型的属性值。当插入或替换文档时，DocumentDB 能够推断文档架构并进行索引，而无需定义架构或辅助索引。

**JSON 文档的树形表示：**



文档的树形表示

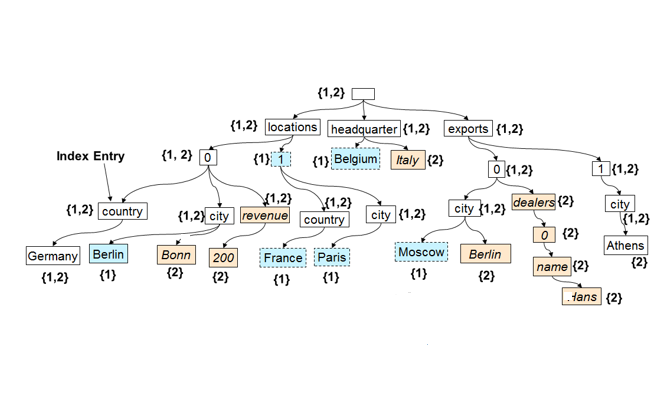
例如，在上图所示的示例中：

* 上述示例中的 JSON 属性 {"headquarters": "Belgium"} 对应于路径/总部/比利时。
* JSON 数组 {"exports": [{"city": “Moscow"}``{"city": Athens"}]} 对应于路径 /exports/[]/city/Moscow 和 /exports/[]/city/Athens。

使用自动索引，(1) 会对文档树中的每个路径进行索引（除非开发人员已将索引策略明确配置为排除某些路径模式）。(2) 每次更新 DocumentDB 集合中的文档时，也会更新索引的结构（即添加或删除节点）。文档自动索引的主要要求之一是确保对深度嵌套结构（如 10 级）的文档进行索引和查询的成本与对包含只有一级键值对的平面 JSON 文档执行该操作的成本相同。因此，规范化的路径表示是建立自动索引和查询子系统的基础。

在路径方面均衡处理架构和实例值的重要意义在于，在逻辑上就像单个文档，如果两个文档的索引在路径和包含该路径的文档 ID 之间保留映射，则该索引也可以表示为一个树。DocumentDB 使用这一事实来生成索引树，通过表示集合中各个文档的所有树的联合来构建。DocumentDB 集合中的索引树会随着新文档增加或集合更新而逐渐长大。

**DocumentDB 索引的树形表示：**



索引的树形表示

虽然是架构自由，但 DocumentDB 的 SQL 和 JavaScript 查询语言仍然可跨文档、空间操作和完全用 JavaScript 编写的 UDF 调用来提供关系投影、筛选器和分层导航。DocumentDB 的查询运行时能够支持这些查询，因为它可以直接对数据的这种索引树表示进行操作。

默认索引策略会自动为所有文档的所有属性编制索引，并提供一致的查询（也就是在文档写入的同时更新索引）。DocumentDB 如何支持索引树的大规模一致更新？ DocumentDB 使用写入优化、无锁以及日志结构的索引维护技术。这意味着 DocumentDB 可以支持持续大量的快速写入，同时仍然保障一致的查询。

DocumentDB 的索引旨在提升存储效率，同时处理多租户。为了追求成本效益，在磁盘上存储索引的开销非常低，并且可预测。同时，可以在为每个 DocumentDB 集合分配的系统资源的预算内执行索引更新。

## 后续步骤

* 下载 [Azure DocumentDB 不限架构的索引”](http://www.vldb.org/pvldb/vol8/p1668-shukla.pdf)，这篇文章于 2015 年 8 月 31 日至 9 月 4 日在超大型数据库第 41 届内部会议发布。
* [使用 DocumentDB SQL 进行查询](/documentation/articles/documentdb-sql-query)
* 在[此处](/documentation/articles/documentdb-indexing-policies)了解如何自定义 DocumentDB 索引