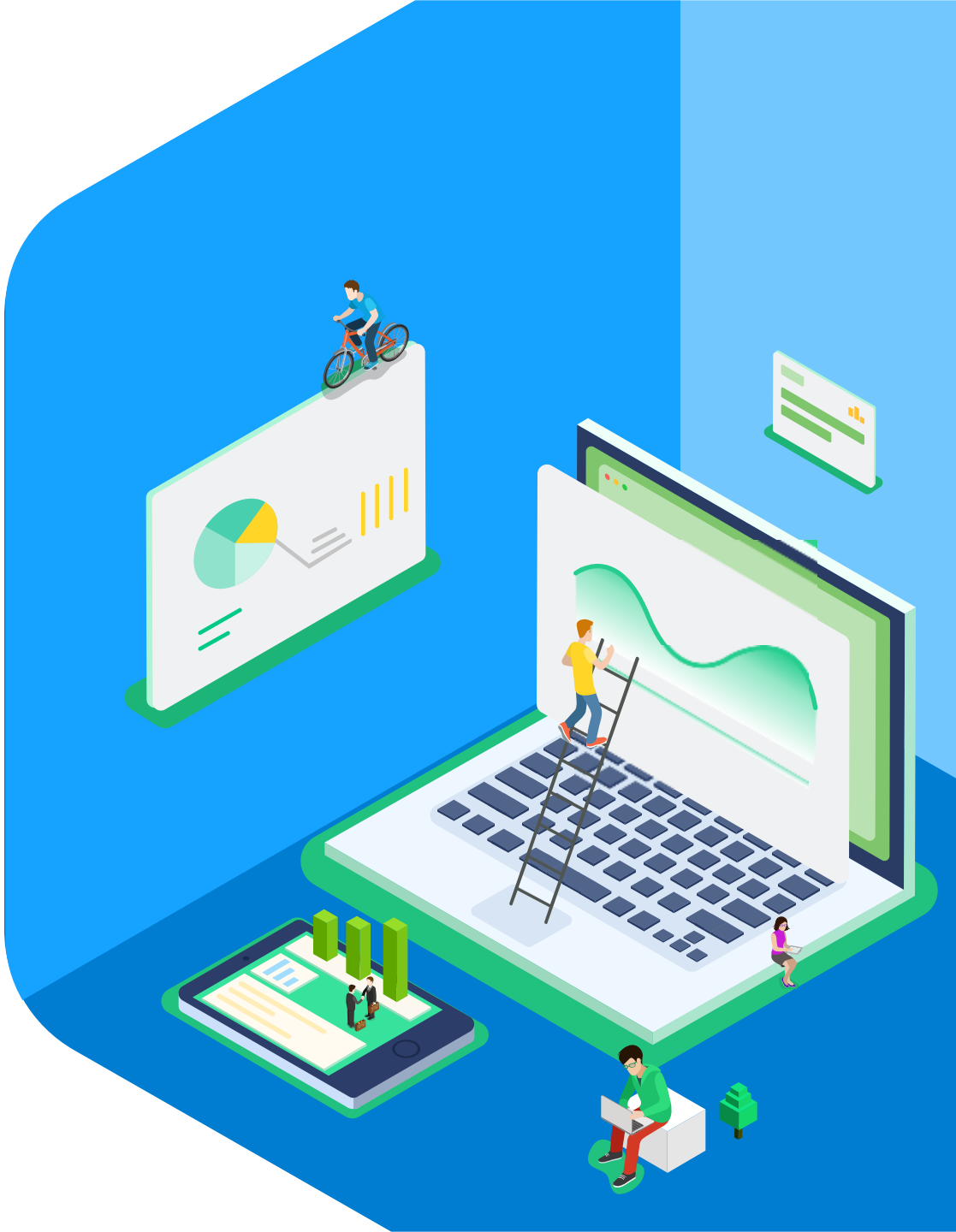
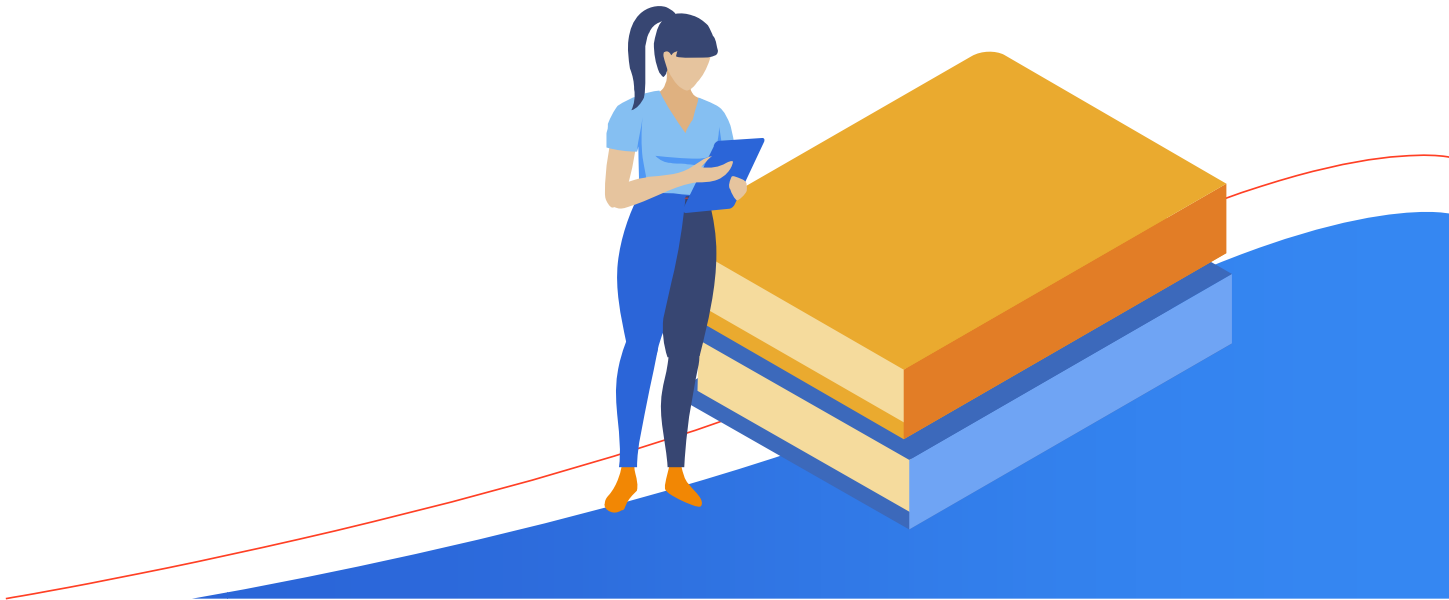
Turing AI 万维图灵 | 大数据系列课程

大数据技术-第七章：HBase数据库

HBase相关知识

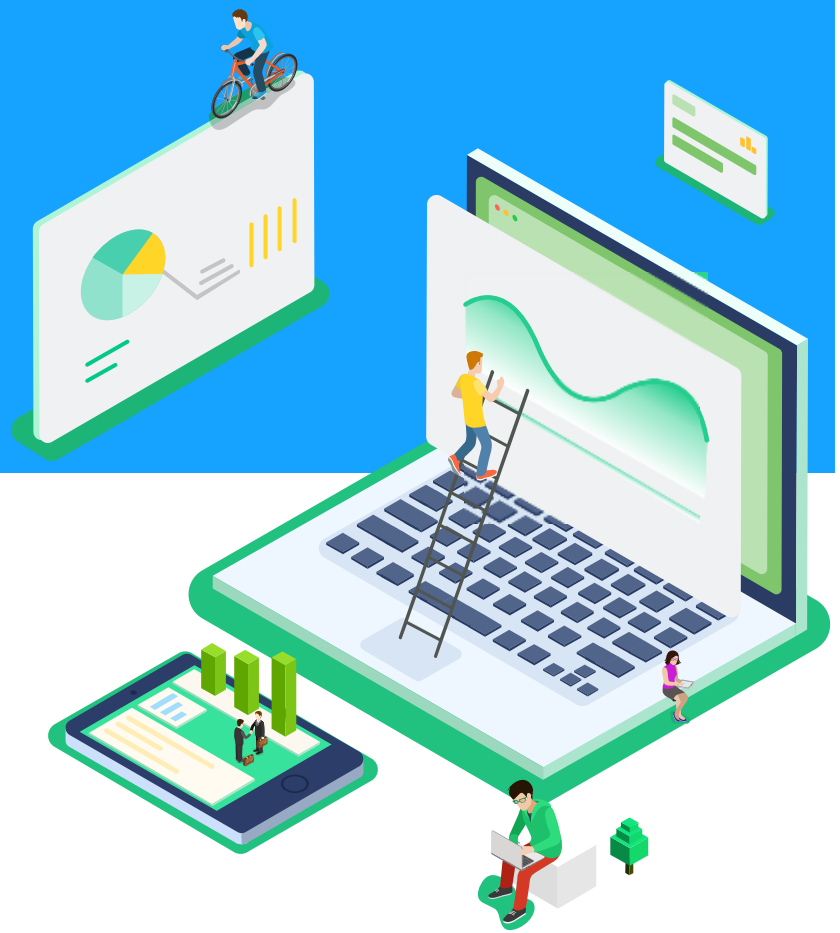


CONTENTS



01

HBase发展历史

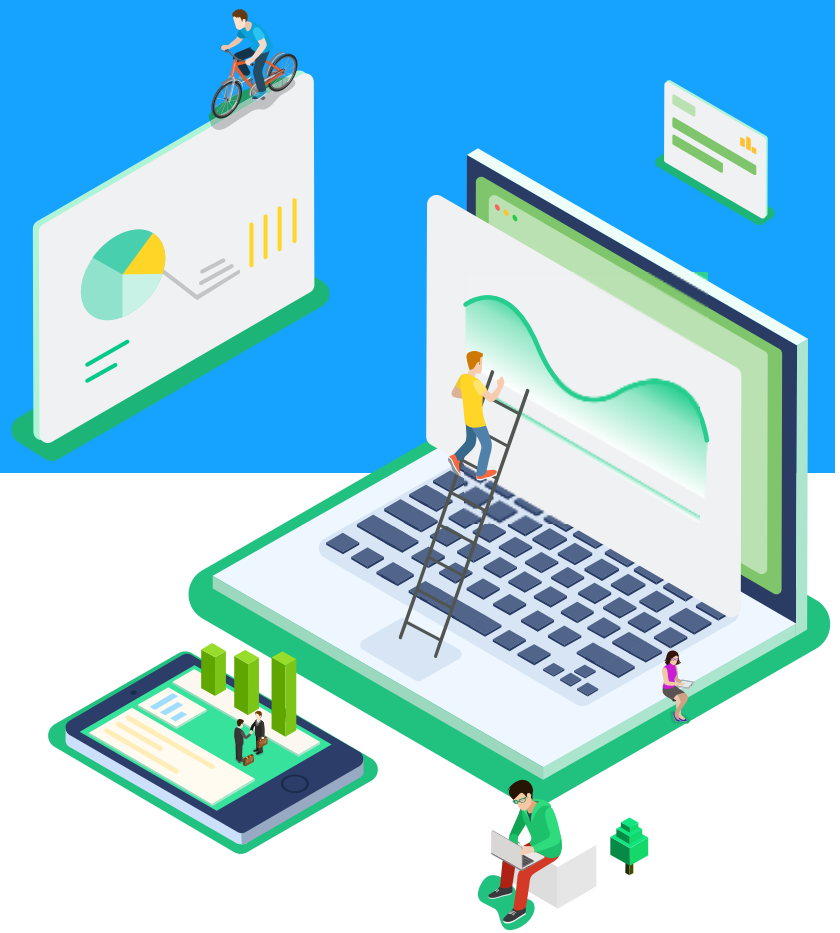


HBase发展历史

HBase（Hadoop Database）是一个高可靠性、高性能、面向列、可伸缩的分布式数据库，典型的NoSQL（Not Only SQL）数据库。它起源于Hadoop的子项目，由Powerset公司在2007年创建，同年10月HBase的第一版与Hadoop 0.15.0捆绑发布，初期的目标是弥补MapReduce在实时操作上的缺失，方便用户可随时操作大规模的数据集。随着大数据NoSQL的流行和迅速发展，在2010年5月，Apache HBase脱离Hadoop，成为Apache基金的顶级项目。次年2011年1月ZooKeeper也脱离Hadoop，成为Apache基金的顶级项目。

02

HBase主要特性

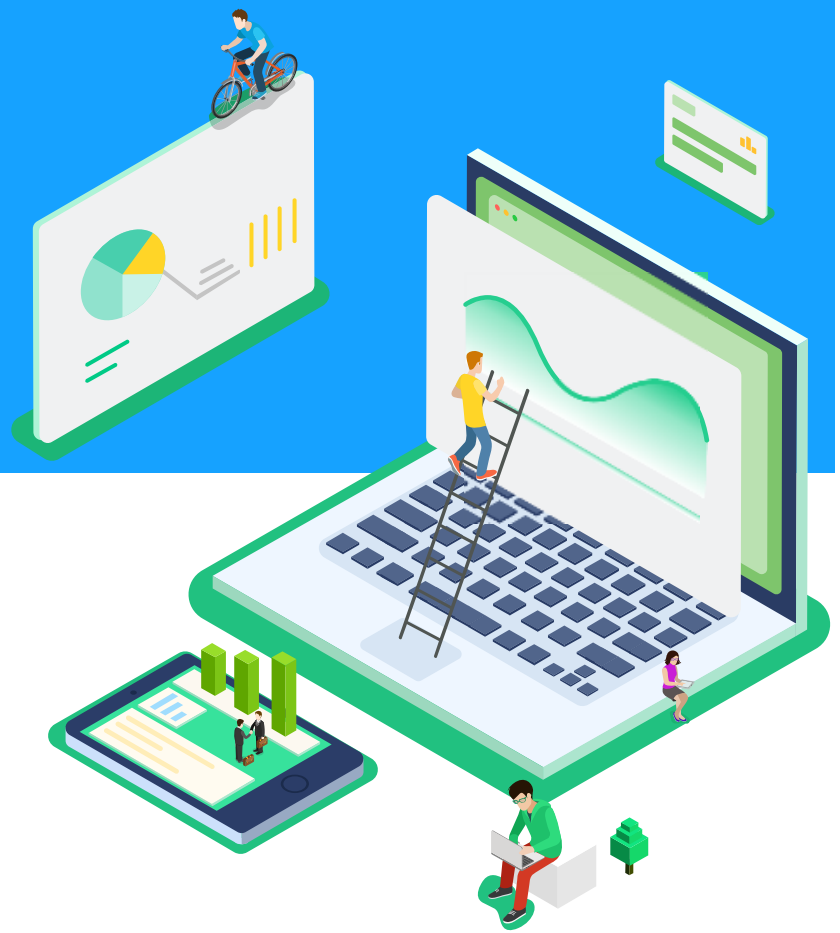


HBase主要特性

- 面向列设计：面向列表（簇）的存储和权限控制，列（簇）独立检索。
- 支持多版本:每个单元中的数据可以有多个版本，默认情况下，版本号可自动分配，版本号就是单元格插入时的时间戳。
- 稀疏性：为空的列不占用存储空间，表可以设计得非常稀疏。
- 高可靠性：WAL机制保证了数据写入时不会因集群异常而导致写入数据丢失，Replication机制保证了在集群出现严重的问题时，数据不会发生丢失或损坏。
- 高性能：底层的据结构和Rowkey有序排列等架构上的独特设计，使得HBase具有非常高的写入性能。通过科学性地设计RowKey可让数据进行合理的Region切分，主键索引和缓存机制使得HBase在海量数据下具备高速的随机读取性能。

03

HBase与RDBMS的区别



HBase与RDBMS的区别

传统的RDBMS具有以下特征：它是面向表格、视图设计的标准化数据，表中的数据类型也会进行预定义，数据保存后表的结构不易修改。每个表格对列的数据有所限制，最大不会超过几个百个，这将导致不同的数据可能会存放到多个表，表格之间存在一对一，一对多，多对一，多对多等复杂关系。正因如此也限制了RDBMS的使用场景更适合于高度结构化的行业，例如医疗，机关，教育等。

HBase是典型的NoSQL代表，它属于一种高效的映射嵌套型弱视图设计，以Key-Value的方式存储数据，每一行数据都可以有不同的列设计。数据依赖于行键作为唯一标识，当行数据的结构发生变生时，HBase也能根据需求做出灵活调整。数据以文本方式保存，HBase把数据的解释任务交给了应用程序，因此它更适合于灵活的数据结构项目。

Turing AI 万维图灵 | 大数据系列课程

大数据

BIG DATA

智 / 能 / 科 / 技　　放 / 眼 / 未 / 来