

# 中英数据库专家谈：数据库的过去、未来和现在

原创

阿里云开发者 于 2021-01-28 14:17:54 发布 577 收藏 1

文章标签：[存储](#) [人工智能](#) [分布式计算](#) [NoSQL](#) [Cloud Native](#) [大数据](#) [关系型数据库](#) [量子技术](#) [数据库](#) [Spark](#)

版权声明：本文为博主原创文章，遵循[CC 4.0 BY-SA](#) 版权协议，转载请附上原文出处链接和本声明。

本文链接：<https://blog.csdn.net/alitech2017/article/details/113336782>

版权

**简介：**数据库是什么？未来的数据会被存在DNA里？数据库里的数据湖是什么？1月16日，扫地僧做了一场直播，请到我的同事——数据库资深专家封神，和来自帝国理工的高级讲师Thomas Heinis（托马斯·海尼斯），2人就数据库这个话题做了比较深入的探讨，老僧印象比较深的是一些前沿的DNA储存大数据等概念。在此老僧奉上双方谈话的全部内容，由于英国学者使用英文讲解，所以对全文进行了中英文的翻译。希望这个速记能帮助对前沿科学有兴趣的同好。



1月16日，扫地僧做了一场直播，请到我的同事——数据库资深专家封神，和来自帝国理工的高级讲师Thomas Heinis（托马斯·海尼斯），2人就数据库这个话题做了比较深入的探讨，老僧印象比较深的是一些前沿的DNA储存大数据等概念。在此老僧奉上双方谈话的全部内容，由于英国学者使用英文讲解，所以对全文进行了中英文的翻译。希望这个速记能帮助对前沿科学有兴趣的同好。

**主持人：**大家好，欢迎来到阿里达摩院扫地僧的直播间，我是扫地僧的小助理。今天，我们的自动驾驶机器人小蛮驴带大家逛了一下阿里云飞天园区，一路没人讲话，不知道盆友们有没有看急了，那接下来我们就聊聊天。

**Moderator:** Hello everyone. Welcome to the live streaming studio of the sweeper monks of Alibaba DAMO Academy. I am the assistant to the sweeper monks. This morning, our autonomous robot Xiaomanlv (Little Competent Donkey) took you guys on a tour in Alibaba Cloud Apsara Park without saying anything. I wonder if you felt anxious or not. Now let's have a good chat.

这次陪我们聊天的是扫地僧的老同事封神。

Today, we have with us our old friend, the Sweeper Monk Fengshen.

**封神（阿里云智能 云原生数据湖分析DLA技术负责人）：**大家好，我叫封神，来自阿里云数据库团队，09年加入阿里，目前主要做数据库数据湖分析方向，主要负责云原生数据湖分析DLA的技术，之前也做了10年左右的数据库与大数据相关的事。

Hello everyone, I am Fengshen from Alibaba Cloud database team. I joined Alibaba in 2009. Currently, I am mainly responsible for Data Lake Analytics, known as DLA. Before joining Alibaba, I had spent 10 years doing things related to databases and big data.

**主持人：**直播间还请到了1位来自远方的客人，Mr. Thomas Heinis，他是帝国理工学院的讲师，请客人自我介绍。

**Moderator:** We are also honored to have with us here Mr. Thomas Heinis, lecturer at Imperial College London. Mr. Heinis, please kindly introduce yourself.

**托马斯·海尼斯（帝国理工数据库专业高级讲师）：**是的，当然。我叫托马斯·海尼斯，现在是帝国理工学院的高级讲师。我在研究小组做研究，我们研究小组基本上负责一切与数据有关的研究工作。我从事大量的数据分析和数据可视化工作，目前也负责数据存储工作，包括所有的新技术，确保我们未来能够有效地分析和理解数据。

Yes, sure. My name is Thomas Heinis. I am a senior lecturer at the moment at the Imperial College in London. I do research here in the research group, which basically takes care of everything to do with data. So I do a lot of data analytics, also, data visualization, and then also data storage at the moment, including all new technologies such that we can basically analyze data efficiently and understand it in the future.



**主持人：**既然请到2位数据库领域的专家，我们今天讨论的主题肯定离不开数据库。数据库对我们这些非专业人士而言，可能最常听的词是删库跑路，而对数据库最直观的理解就是Excel表格。先给我们的观众介绍一下什么是数据库。请问Mr. Heinis，你给学生上第一堂课时如何介绍？

**Moderator:** Since we have invited two experts in the field of databases, the topic of our discussion today is certainly inseparable from databases. For non-professionals like us, probably the most common words we have heard is dropping a database. And our most intuitive understanding of databases is Excel tables. First, please introduce to our audience what a database is. Mr. Heinis, how would you explain databases to freshmen in your first class?

**托马斯·海尼斯:** 我通常会给学生稍微介绍一下数据的由来, 这与银行密不可分。上个世纪六七十年代期间, 银行储存了大量的数据。它们需要将这些数据组织起来, 关系型数据库应运而生。数据库本质上是很多信息和数据的集合, 这些数据被组织起来, 从而实现高效的分析、访问、管理和更新的目的。所以, 计算机数据库通常来自于数据文件的导航, 从传统上而言, 或者就历史渊源而言, 数据库确实来自银行, 有很多关于银行客户的信息以及他们账户余额的信息。

I usually explain a little bit of history about it. That's all to do with banks. Banks had a lot of data in the 60s and 70s. And they needed to organize that data. That's kind of where relational databases come from.

Essentially, what I tell students is that a database is a collection of lots of information, lots of data that are organized so that it can be analyzed, accessed, managed, updated very efficiently, right?

So computer databases, typically from the navigation of data files. Databases, traditionally, or historically, do come from banks. There was a lot about bank customers and clients their balance of their accounts and information.

And within databases, then we have this massive for this big branch of this big technology called relational database, which is really where we organize data to tables, rows, columns, which really contain information about customers, clients, transactions, sales, etc. and all kinds of very well structured information.

而在数据库里面, 有一个庞大的技术分支, 叫做“关系型数据库”, 其实就是把客户、交易、销售等信息以高度结构化的形式组织到表、行、列当中。而有些同学一开始就已经听说过SQL, 也就是用于查询数据库的查询语言。最近几年, 情况有所变化。但总的来说, 关系型数据库确实是在上世纪六七十年代期间由银行的应用案例所驱动的。过去二十年间, 情况发生了巨大的变化。因为我们有了需要组织数据的新型应用, 比如科学应用, 或者是社交网络等类型的应用。

基本上, 这些应用需要略微不同的数据库。因此, 后来我们转向了NoSQL数据库或非关系型数据库, 有了不同的用例, 也开始管理更多的数据, 也就是现在所说的大数据。简单地说, 我们收集了海量的数据, 需要用数据库来分析和存储这些数据。

And some of the students have already, on the start, already heard about SQL, which is the query language to ask for a query database.

In recent years, things have changed a little bit. So relational databases, they do come from the 60s 70s driven by this banking use case, for this banking application.

And in the last 20 years, things have changed drastically, right?

Because we have new applications that need to organize the data such as scientific applications, or the types of applications like social networks, and etc.

Basically, these applications require slightly different databases. And so that's where then we went to kind of noSQL databases or non-relational databases and we moved to different use cases. And we also move to managing much much more data, What is nowadays called Big Data. Basically, we collect tremendous amounts of data. And then we need to come to databases to analyze and store these data.

**主持人:** 想问从事这个行业的封神老师, 进入到现代, 数据库为何越来越重要?

**Moderator:** I have a question to Fengshen. Why are databases more and more important in the modern era?

其实数据库一直很重要, 最为简单的一条就是 数据不能丢。企业如果丢了最为核心的数据库, 则企业可能直接面临破产。

Actually, databases have always been important. To put it simply, data cannot be lost. If an enterprise loses its core database, it might directly go bankrupt.

为什么数据库越来越重要 那是因为数据还蕴藏着宝藏，之前数据存着也就存着，一般存核心的数据，如交易、客户、商品的数据，一些日志数据存就是为了解决查询问题。也不会去埋点，更加谈不上去爬取或者购买数据了。

Why are databases more and more important? That is because data also contain treasure, In the past, data were just stored there. Generally, core data were stored, such as the data of transactions, customers, and goods. Some log data were stored just for query purposes. There were no buried points, let alone crawling or purchasing data.

互联网也经历了好几个阶段，从开始的新闻门户时代，到可以有交互的类似BBS、淘宝购物时代，到all in无线，到智能时代，产生的数据也越来越多。

The Internet has gone through several stages, from the news portal era in the beginning, to the era of interactions with such platforms as BBS and Taobao, to the era of wireless technology, and then to the era of intelligence. More and more data have been generated in the process of evolution.

-从数据量来看，IDC统计，2005年全球数据是130EB，2019年为41ZB，涨了322倍；

-According to the statistics released by IDC, the amount of global data generated in 2005 were 130 EB, and that in 2019 was 41ZB, an increase of 322 times.

-从数据应用来看，越来越多的公司也使用数据做出了出色的成绩，阿里、头条、百度、滴滴等Top100知名互联网企业都是数据驱动的企业。我们再来看传统的产业，有智慧园区、城市大脑、县域大脑、智慧农业、智慧城市、智慧医疗、工业4.0等等，也都在使用数据技术在赋能各个产品，帮助这些产业数字化转型，提升效率。在产业应用看，要充分发挥想象的空间，使用数据赋能产业转型成长。

-From the perspective of data application, more and more companies have also made outstanding achievements with data. For example, the Top 100 well-known internet enterprises, including Alibaba, Toutiao, Baidu and DiDi, are all data-driven enterprises. In terms of the traditional industries, data technology is used in smart parks, city brain, agriculture brain, smart agriculture, smart cities, smart medicine and Industry 4.0 to empower each product and help the respective industries achieve digital transformation and increased efficiency. Seen from the perspective of industrial application, it is necessary to give full play to imagination and use data to empower industrial transformation and growth.

-从国内形势看，国家也提出了新基建，核心是以云计算、大数据、人工智能、5G、区块链为核心，这些核心中的核心是数据的应用，再过10年，真的是万物互联的时代，数据量增长的速度会更加快；在疫情时代，有研究机构研究表明，疫情让数字化转型快了5年左右。

-Seen from the domestic situation, the Chinese government has also put forward the concept of new infrastructure. Its core technologies include cloud computing, big data, artificial intelligence, 5G and blockchain. And the core of these core technologies is the application of data. It is expected that in 10 years, we will enter the era of the Internet of Everything, when the growth of data volume will be even faster. Relevant institutions have found through research that the COVID-19 pandemic has expedited the digital transformation by 5 years.

-从高校和研究机构看，国内高校专业增设最多是大数据技术、人工智能的专业；总之，21世纪除了人才，还有什么最贵，那就是数据，数据相当于20实际的石油，是21世纪整个社会效能运转的润滑剂。因为数据越来越重要，所以数据库越来越重要，数据库是这一切的核心载体。

-From the perspective of universities and research institutions, big data technology and artificial intelligence programs have been widely offered these days. In short, in addition to talents in the 21st century, what else is the most expensive? The answer is data. Data are equivalent to the oil of the 20th century. Data are the lubricant for the functioning of the whole society in the 21st Century. Due to the increasing importance of data, databases are also becoming more and more important. And databases are the core carriers of everything.

主持人：中国和英国的对大数据的定义有什么不同？这会导致双方程序员对数据库的理解不同吗？

**Questions:** What is the difference in the definition of big data between China and the U.K.? Will such differences lead to different perceptions of databases between the programmers in the two countries?

**封神：**大数据其实我认为没有准确的定义。比如，如果数据量比较小，但是训练用的机器比较多，也可以认为是用到了大数据技术。我一般认为用数据驱动业务发展就属于使用了大数据相关的技术。之前国内提Big Data（中文指：大数据）比较多，现在国外提Data Lake（中文指：数据湖）的概念比较多，主要还是云公司在主导，数据很多存在了对象存储上。阿里数据库团队提的是库、仓(Data Warehouse)、湖(Data Lake)、多模（Multi-Model），并且我们还专门做了一个云原生数据湖分析DLA的产品。另外，我们看到著名的咨询公司Gartner把大数据报告合并到了数据库报告里面。

**Fengshen:** I don't think there is a precise definition of big data. For example, if the amount of data is small, but many machines are used for training, it can still be considered that big data technology has been used. I generally think that adopting the data-driven business mode is equivalent to using big data-related technologies. In the past, the concept of Big Data was mentioned a lot in China. Nowadays, the concept of Data Lake becomes popular in foreign countries. In most cases, cloud companies are taking the lead. And a lot of data exist in object-based storage. Alibaba's database team mentions data library, Data Warehouse, Data Lake, and multi-model. We have also specifically made a product based on Data Lake Analytics (DLA). In addition, the renowned advisory company Gartner Inc. integrated big data reports into its database report.

我认为数据库包括传统的数据库技术（如MySQL、PG）；也包括数据仓库、数据湖的技术，如开源的Spark、Hadoop，阿里的ADB、DLA等，也包括最近流行的LakeHouse技术。

In my view, databases include traditional database technologies (such as MySQL and PG) as well as data warehouse and data lake technologies, such as open-source Spark, Hadoop, Alibaba's ADB and DLA. They also include the Lakehouse technology which has been quite popular these days.

我跟英国的程序员交流比较少，跟北美有一些交流，整体应该理解差不多。技术传播的速度也比较快，大家理解应该比较类似。由于中国的市场比较大，由于某一些原因，数据也比较多，这些会加快对数据的应用的发展。

I haven't had much communication with programmers in the UK. But I have had some communication with programmers in North America. Overall, our perceptions are more or less similar. The speed of technology communication is also relatively fast. So, the understanding should be more or less similar. Besides, the huge market in China and the larger amount of data generated here will accelerate the development of data application.

**托马斯·海尼斯：** On some level, yes. I think this is gonna be long long rounds. But I don't think you know, if you look at it, from a technical perspective, the types of data could be the same, it's going to be very similar. But what I do think is that the scale is massively different. And that's quite, you know, just because there's so much more data available in China.

某种程度上，会的。这一点说来话长。但我认为，从技术角度而言，数据的类型可能相同，或者非常相似。但数据的规模却差异巨大，这主要是因为在中国，可获取的数据更多。

And it's because I think personally, I think it's because the society is a little bit more technologically advanced, or is it easier to... I think China adopts technology easier, which means that, for example, you have more sensors everywhere with traffic measurements, and an engine mentioned the DiDi, right, which, which, you know, produces the same kind of data as Uber, for example, but on a different scale, right. And that applies to everything.

我个人认为，这是因为相比之下，中国社会的科技更为发达，并且中国更易采用技术。这就意味着，例如，中国的路面上会有更多的传感器来监测交通情况。又如，滴滴打车软件所产生的数据和Uber所产生的数据一样，但是规模却不同。其他领域也是如此。

And what he said about the pandemic, being a catalyst for this transformation is absolutely also absolutely true. Like we have made way more electronic payments, now, we have just everything is more digital, and I think a lot of it will stay digital, and that produces data that produces data that we need to analyze.

刚刚他提到说新冠肺炎疫情是推动这一转型的催化剂，这一点毋庸置疑。比如，我们现在使用电子支付的频率比以前高了很多，生活的方方面面都更加数字化了，我认为数字化的趋势会持续下去，这就产生了需要我们分析的数据。

So with what I mentioned, about China, being a technological, a bit more about having more technical technological affinity means that we basically have more places where we collect data, and much more sensors, people interacting online that produces more data.

我刚提到说中国更亲近技术，这就意味着在中国，数据的来源更多，有更多的传感器，网络用户互动所产生的数据也更多。

And then also, you know, China's population is huge. So that also means that more data is being produced. So it's, I would say, you know, there are two technical challenges when it comes to big data or databases in general.

此外，中国人口众多，因而产生的数据也更多。因此，我认为，大数据或数据库主要面临两大技术挑战。

One is the data formats, you know, that's life changing, as mentioned, is going towards also data lake right loads of different formats. But I don't think that differs much between China and the rest of the world.

一是数据格式上的挑战，其发展趋势是走向数据湖等不同类型的格式。但在这一点上，中国和世界其他地方的区别不大。

But what what is really different is the amounts of data. And so that means that we need whatever we develop the analysis, visualization, storing the data that needs to happen efficiently on a much, much larger scale, which then again, brings in a lot of technical challenges and challenges as well. Right. So I think i think that's that's, that's what I think that's the difference. But I think the definition as such, it's roughly roughly the same.

中外真正的区别在于数据量。具体而言，中国需要更大规模和更高效地进行数据分析、可视化和存储数据等任务。这又带来了许多技术挑战。我认为中外在数据库领域的区别就在于此，但两者对数据库的定义基本相同。

It's also I also have the feeling That, that, you know, China has a different, the Chinese people have a different understanding of personal data as well,

此外，我认为中国人对个人资料的理解与外国也不同。

it's very, very difficult for us to get data from companies here or from you know, there's always a perception of these breaches data privacy, whereas in China and appear of works a lot with with Chinese University as well. So we get a lot of data from I don't think it's DiDi, but somebody saw some similar data sets, it's quite easy. So there's also kind of like, I feel like China has really this, this kind of this, this more affinity towards technology, and this is a and like, this more of a project, let's see what we can do with the data. Let's see if we can improve things, you know, so we're also collaborating on a traffic optimization project that you know, they collect massive amounts of data about which vehicle passes through the road, where at what speed what's the congestion level? Can we remove traffic and all these kinds- of things? It's really kind of like a very pragmatic approach to using data really, what can we do to improve you know, everything really. That's what data does these days.

在英国或者其他国家，从公司获取数据非常困难，因为外国企业总是担心侵犯数据隐私。相对而言，从中国企业或者院校获取数据容易得多。因此，我们获取了很多数据，可能不是直接来自滴滴公司，但数据集也比较类似。总地来说，在中国，数据的获取更为容易。我感觉中国更亲近技术，更愿意使用技术，更愿意利用技术来改善现状。我们正在与中方合作一个交通优化的项目，他们收集了大量的数据，包括经过道路的车辆信息，车辆的位置，车辆的行驶速度，路面的拥堵程度，是否可以消除拥堵等。这种数据使用真的非常务实，相当于使用数据来改善一切可改善之处。这便是数据在当今社会所发挥的作用。

主持人：随着大数据这个概念的出现，数据库是怎么进化的，请封神老师讲讲？

With the emergence of the concept of big data, how has databases evolved over the years? Fengshen, please share with us your view on this question.

封神：数据库怎么进化，先看看数据库怎么来的。从广义看，数据记录的历史早就有了。在5000年前，人类开始用绳结计数；在2000年前有纸张，到1946第一台计算机的诞生；计算机的诞生后，才有了现代意义的数据库。为了形象说数据库是什么，好比 你有一个管家，管家有一个记账本，你每天花费多少钱，收入多少都会告知管家，管家记录下来。你就可以知道你目前多少钱，每个月花费多少钱；数据库就是管家+账本；管家提供计算力，账本提供存储；

To figure out how databases have evolved, let's first look at how they came into being. In a broad sense, data recording dates back to a long time ago. Back 5000 years ago, humans began to count with knots; 2000 years ago, paper started to be used. And in 1946, the first computer was invented. After the invention of the first computer, the databases in the modern sense came into being. Let me use an analogy to explain what a database is. Suppose you have a housekeeper, and the housekeeper has a ledger. You inform the housekeeper how much you spend and how much you earn every day. The housekeeper records your income and expenditure in the ledger accordingly. You can then know how much money you currently have and how much you spend each month. In this case, the database is the housekeeper + the ledger: The housekeeper provides the computing power, while the ledger provides the storage.

数据库也发展经历了很多阶段，为了“记好账”，数据库也在不断演进。我们一般根据把数据库发展分为了4个阶段：

The development of databases has gone through several stages. And in order to “keep good accounts”, databases have also been evolving. We generally divide the development of databases into 4 stages:

1970~1990商业数据库时代 收费时代

The two decades from 1970 to 1990 was an era of business databases when fees were charged for the use of databases.

1990~2000 开源数据库时代 开源时代

The decade from 1990 to 2000 was an era of open-source databases.

2000~2015 互联网浪潮 大数据时代（大数据计算、存储、NoSQL）

The period from 2000 to 2015 was an era of Internet and big data (big data computing, storage, NoSQL)

2015~现在 云的浪潮 云原生时代+AI

Finally, the era from 2015 to the present is an era of cloud technology and the widespread application of Cloud Native and AI technologies.

数据库的发展跟几个因素有关，硬件的发展，需求；硬件主要指 存储、网络、内存、CPU。存储就是存数据，内存与CPU关系到计算力，网络就是传输。

The development of databases is related to several factors, including the development of hardware and the market demand; hardware mainly refers to storage, network, memory and CPU; storage refers to data storage; memory and CPU are related to computing power; and network concerns transmission.

大数据这个词语大概在10年前开始流行，大数据系统开始独立于数据库系统发展的，随着最近5年的发展，大数据相关技术又慢慢与数据库技术结合回归到数据库的大家庭。比如，2020年，著名的咨询公司Gartner把大数据报告合并到了数据库报告里面。最为典型的是 DalaLake的发展，融入了事务&MVCC的概念，NewSQL的发展，NewSQL也融合了分布式的理论，并且还有一个HTAP的方向在探索。目前数据库领域分为TP、NoSQL、AP等领域。TP一般有单机、分布式、事务型的数据库；NoSQL就相对散一些：宽表、图、文档、时序、时空等；AP有Data Warehouse、DalaLake领域。

The term big data became popular around 10 years ago, when big data systems started to develop independent from database systems. With the development in the last 5 years, big data-related technologies have slowly returned to the database family by combining with database technologies. For example, in 2020, the famous research and advisory company Gartner integrated its big data report into its database report. And the most typical case is the development of Alibaba's Dala Lake Analytics team, which incorporates the concepts of transactional databases & Multi-Version Concurrency Control (MVCC). Besides, the development of NewSQL also incorporates the theory of distributed databases. And the direction of HTAP is under exploration. Currently, the database field consists of such segments as TP, NoSQL and AP, TP generally consists of standalone databases, distributed databases and transactional databases; NoSQL covers a relatively wider scope, including wide tables, graphs, documents, time sequence and space-time. AP consists of such fields as data warehouses and data lakes.

在企业界，肯定是做看得见，并且在5年内能落地的事情。未来5年，数据库领域核心发展方向是云原生+分布式，具体讲：Serverless、数据库与大数据一体化、智能化、安全可信、软硬件一体化、离在线一体化、多模数据处理。举个例子，我负责做的云原生数据湖分析DLA就是传统大数据、Hadoop、Spark的升级，需要融合传统数据库技术，并且基于存储与计算完全分离的云原生架构。我们选用对象存储，支持常见的消息、TP和NoSQL数据库系统数据的归档。我们一般归档到DataLake里面。还支持了一些事务、版本的东西，并且把Spark、Presto等组件做成云原生的弹性、随时可用，即开即用，按需计费，分离后带宽的损耗通过引入本地的Cache解决。

The prospects of the business community in the next 5 years are definitely foreseeable. In the next 5 years, the core development directions of the database field are Cloud Native and distributed databases, which specifically include serverless, integration of databases and big data, intelligence, security and trustworthiness, hardware and software integration, offline and online integration, and multi-mode data processing. For example, I am currently responsible for Data Lake Analytics (DLA), which can be considered the upgrade of the traditional technologies such as big data, Hadoop and Spark. It requires the integration of traditional database technologies and is based on the Cloud Native architecture of complete separation of storage and computing. It selects object-based storage and supports the archiving of common messages as well as the data in TP & NoSQL databases systems. Normally we archive the data in the data lakes. Besides, DLA also supports transactions and versions. Besides, the Spark and Presto components are also incorporated to achieve elasticity for the Cloud Native, which is accessible at any time and charged on the basis of demand. The loss of bandwidth after separation is solved by introducing local Cache.

**托马斯·海尼斯：**其实封神的分享已经很详细，并无太多可补充之处。如果非要补充的话，如您方才所言，开源数据库大大推动了数据在社区内的扩散。我认为相较以前，数据库的使用也变得容易得多了。

Well, there's not much more to add to what Fengshen has already said. Right. But I think I think what would what would I do want to maybe add really is that also, you know, Like you said absolutely correctly, is that open source databases have done a tremendous service to the community in kind of getting databases everywhere. I will say that I think it's also kind of the databases that have become much, much, much easier to use as well. And, you know, years back the first year students, they'd never seen a database. Today, if I asked, they've all seen MongoDB, or other technologies, kind of like easy-to-use databases, you know, not relational databases necessarily, but easy to use technology.

比方说，若干年前，大一的学生从未见过数据库。而现在的大一新生。几乎都见过MongoDB或者其他一些易于使用的数据技术。他们未必见过关系型数据库，但基本都见过易于使用的数据科技。

And that really has made a difference in terms of training people, they also kind of has changed their understanding the approach of people to using databases. Now, back in the day, people were storing data just in RAW files. Nowadays, they know, if I want to have efficient access to the data, then I need to use a database and they know how to do so.

这对人员的培训起到了巨大的助推作用。这也改变了人们对数据库使用方法的理解。以前，人们只是把数据存储于原始档案中。现在，他们知道，需要使用数据库才能有效地访问数据，并且他们也知道如何使用数据库。

So databases have really changed, or kind of database have become much more pervasive. They're used everywhere these days. So that has definitely changed.

所以，数据库真的发生了变化，或者说数据库的使用变得更加普遍，现在各地的人们都在使用数据库，这是一个很明显的变化。

And now I unfortunately, forgot your question, which I didn't answer.

不好意思，我忘记你提的问题了，所以没有回答。

So essentially, what really changed, right, and I said this initially, already kind of databases were designed relational databases were designed for banking applications that were revolving around transactions, which was really the centerpiece of banking applications. And that has made a lot of design decisions difficult.

至于数据库发生了什么样的演变。之前我有提到，数据库，或者说关系型数据库是为银行的应用而设计的，是围绕交易设计的，而交易是银行应用的真正核心。这也使得许多设计决策变得困难。



And then in recent years, like XX had mentioned, right, databases are kind of new use cases emerged, all of a sudden, we no longer have the data we have along fits nicely in a in a table, we actually have a graph.

然而，最近几年，正如封神刚提到的，出现了一些数据库的新用例。突然间，我们获得的已经不再是表格数据，而是图形数据。

So we have graph databases, or we realized a lot of data is natively very structured in a document, XML or similar. So we develop document databases.

于是我们有了图形数据库，或者说我们意识到很多数据天生就是高度结构化的，类似于文档数据或XML数据。所以我们开发了文档数据库。

So there's, there's the now we, in the early, maybe around 2000, a little bit after 2000, we had this understanding that one size doesn't fit all. So we need to have different types of databases. So I mentioned graph database, document databases. But there have also been other other databases, very customized databases. For scientific applications, right?

所以，大约在2000年左右，具体说是刚过2000年的时候，我们意识到不能一刀切，而是需要不同类型的数据库。我刚提到了图形数据库和文档数据库。但是也有其他类型的数据库，比如用于科学应用的高度定制化的数据库。

They produce massive amounts of data, like physics experiments, like astronomy, DNA experiments in biological experiments, they have all kinds of their own database technology these days,

这些科学应用数据库会产生大量的数据，比如物理实验、天文学、生物实验中的DNA实验等。这些领域都有各自不同类型的数据库技术。

Back in the day, we've tried to fit everything in relational database and it didn't work really well. So each one of those now as their own title type of database.

以前，我们试图把所有数据都装进关系型数据库，但效果并不理想。因此，现在不同领域都有各自不同类型的数据库。

At the same time, we also, you know, more and more data has been produced in different formats. And this is really where the kind of what is this notion of a data lake of engine has been mentioning is coming from that we have tons of data in different formats, we still want to analyze the data as a whole. So we need some sort of kind of some sort of integration between that or some sort of way of analyzing heterogeneous different data types. And that has also changed. So we have now this capability to just produce data, throw it in a database, Put simply, and then analyze it efficient, efficiently, efficiently at scale. Right. So that's really how things I believe have changed.

同时，越来越多数据也在以不同的格式产生，因此我们不断提到数据湖引擎的概念。之所以引出这个概念，是因为我们有大量的不同格式的数据，但仍然希望将数据作为一个整体来分析。所以我们需要某种整合技术，或者说需要采用某种方式来分析不同类型的数据。所以我们现在有这样的能力：即把数据生产出来，简单地扔到数据库里面，然后高效地进行大规模的分析。我认为这就是数据库所发生的演变。

There's also other trends like cloud computing, in general, which has made it which is also supported for particularly smaller businesses to have their own database their own data solution. Because they no longer need to own the resources. And they can just if they have a big analysis to run on their data, or they just use cloud resources to do so temporarily, without having the hassle of owning. Right. So that has also held,

还有其他的趋势，比如云计算。总的来说，云计算帮助小微企业拥有自己的数据库和自己的数据解决方案，因为它们不再需要拥有资源，只需要掌握数据的分析能力，或者只是暂时使用云端资源来进行分析，而不需要拥有资源。所以云计算对小微企业起到了助推作用。

then we also have a huge trend in terms of hardware. So we have, obviously we have better hardware. And every now and again, the database community tries to really optimize the database for new hardware beat is multi core processors, which are not particularly new, but all kinds of hardware aspects of new CPUs, new types of memory, non volatile memory, for example, change a little bit how we organize and analyze data. So a lot of hardware trends has also changed or shaped database, database technology.

此外，硬件方面也有很大的发展趋势。显然，我们有更好的硬件。而且每隔一段时间，数据库社区就会尝试真正的优化数据库，针对新的硬件采用多核处理器，这也并不新奇，但是硬件各个方面的优化，比如使用新的中央处理器、新的储存器，非易失性存储器等，在一定程度上改变了我们组织和分析数据的方式。所以说，硬件的很多发展趋势也改变或者塑造了数据库技术。

And then finally, what has happened in the last couple of years is really the use of machine learning or artificial intelligence in and around data and that has driven a lot of research and has also produced a lot of products. And when I talk about AI or artificial intelligence databases, it's really kind of The database research community has taken an approach And has done a lot of different things

最后，过去几年间，机器学习或人工智能技术开始应用于数据领域或与数据相关的领域，这推动了很多研究，也催生了很多产品。谈到人工智能或人工智能数据库，数据库研究社区做了很多不同的事情。

for example, you know, artificial intelligence machine learning requires a lot of learning, which requires a lot of data, and for that we need to have data that is clean and has been processed and has been manual has been brought in the right format. So, that's a database task.

例如，人工智能和机器学习的发展需要大量的学习，这就需要大量的数据。为此，我们需要有干净的数据，经过处理的数据，和经人工处理为正确格式的的数据。这是数据库层面的任务。

And then the learning itself is also to some degree a database. Right And so, we have worked on that, that has had a tremendous impact in recent years.

此外，从一定程度上而言，学习本身也是一个数据库。我们在这方面也下了不少功夫，这在近几年产生了巨大的影响。

We also use artificial intelligence within the database itself to accelerate the database accelerate query execution the analysis. And then we also use artificial intelligence to organize the database itself.

我们还在数据库内部使用了人工智能技术来加速数据库的索引、执行和分析。我们也使用了人工智能技术来组织数据库本身。

so that so I would say that artificial intelligence is a mega trend of course, we all know and you know has touched all aspects of life but it is also interesting enough to touch databases which not just touched but changed profoundly how we design and use databases.

因此，我认为，人工智能当然是大趋势。众所周知，人工智能已经触及到我们生活的方方面面，但更有趣的是，它也触及到了数据库领域。准确的说，不仅是触及，而且深刻地改变了我们设计和使用数据库的方式。

**主持人：**刚刚两位老师说到了很多关于数据库的基础知识，如果我现在给这场直播起个名字，我会叫它“数据库入门必看”。开玩笑，我们实际上是个前沿学术分享的直播。其实任何一门学科在学界和工业界都有2种形态，在工业界落地很重要，你能在工业界为一门学科找到很多应用场景，比如刚刚封神老师讲到的双十一、工业大脑，而学界的探索往往非常有想象力。我们很想请2位来展望一下，数据库的未来会如何发展？比如5年内、10年内、50年内、100年内？

Our two teachers just shared a lot of basics about databases with us. If I were to give this livestreaming interview a name, I would call it "Database Essentials". Just kidding. This is actually a live interview about cutting-edge database technologies. In fact, any discipline exists in two different forms, one in academia and the other in industry. It is important to apply technologies in industry. You can find many application scenarios for a discipline in industry, such as Double 11 Shopping Festival and Industrial Brain mentioned by Fengshen just now. In comparison, the exploration in academia is often very imaginative. We would love to ask you two to look ahead into the development of databases in the future. For example, what will databases be like in 5 years, 10 years, 50 years, or even 100 years?

**封神：**我关注的一些方向，未来5年，数据库领域核心发展方向是云原生+分布式，具体讲：Serverless、数据库与大数据一体化、智能化、安全可信、软硬件一体化、离在线一体化、多模数据处理，这个会对每个数据库的每个子领域都有影响。具体在学术界研究的，我看的还相对模糊一些。按照人类发展来看，发展应该是越来越快。不过，计算机还是冯诺依曼架构，未什么时候会颠覆，目前我也没有概念。10年是什么样，我其实压根不知道。目前唯一的的就是保持敬畏之心，保持学习。

I would like to talk about some of the directions I focus on. In the next 5 years, the core development directions of databases would be Cloud Native and distributed databases. Specifically, I'm talking about serverless, integration of databases and big data, intelligence, security and trustworthiness, software and hardware integration, offline and online integration, and multi-mode data processing. These technologies will have an impact on each subfield of each database. As to database research in academia, I only have a vague idea. Seen from the history of human development, the development of databases should be faster and faster. However, computers nowadays are still based on the von Neumann architecture. I have no idea when it will be replaced. And I actually have no idea what kind of development will have happened in 10 years. At present, the only thing I am sure about is to maintain a sense of awe and keep learning.

主持人：学术界就是Mr.Heinis的研究方向了，请 Mr.Heinis 继续来说

托马斯·海尼斯：Yeah, well, what's the future? It's difficult to predict, right? But in terms of, you know, kind of like a five year perspective, the only thing I would add in that I think will make a difference in the in the short term is probably also, like I mentioned, AI, artificial intelligence helping us a little bit to organize the data to accelerate analysis, etc.

未来会如何？我们很难预测。但就未来5年而言，短期内可能出现的进展，就是我刚提到的：人工智能将在一定程度上帮助我们组织数据和加速数据分析。

So I think we're kind of lucky that this is the case. Because a lot of students want to work on AI. And if you can kind of combine as a database technology, we get a lot of talented students involved. But yeah, so I think in the short term, I think AI will also have an impact on databases.

从这点来看，我认为我们很幸运。因为很多学生想研究人工智能。如果能把数据库技术结合起来，我们就能吸引大量人才参与进来。所以，我认为在短期内，人工智能也会对数据库产生影响。

I think also that visualization will become important. And we move there to virtual reality, right, kind of which, which offers us a much more, much more kind of, you know,

可视化技术也会变得更加重要，以及与之相关的虚拟现实技术。

we can kind of interact with the data, we can touch the data to some degree, you know,

它能帮助我们与数据互动，在某种程度上，我们将能“触摸”到数据。

I used to do research and have a feed with gloves with haptic feedback, we can touch to data, this kind of thing will I think will become more important not for an individual analyzing data. But I think for collaborative analysis of data to analyze data together to understand it together, I think that's where we also need to put in and put some research to kind of like help to, to find easier ways for people to understand the impact of things.

我曾做过相关研究，戴上触觉反馈手套，我们可以“触摸”到数据。我认为这类技术会变得更加重要，不是针对个人分析数据而言，而是针对数据协同分析，即团队共同分析和理解数据。这也是我们需要投入研究的地方，以便找到更简单的方法，帮助公众理解数据及其带来的影响。

And then, like Fengshen said right, at one, one important thing that's going to happen fairly soon, probably five to 10 years, maybe a little bit more, it's going to be quantum

另外，正如封神所言，不久的将来数字领域将发生重大突破，那便是量子科技。这也许会发生在5到10年之后，也许更久一点。

and quantum. You know, it's difficult to fathom what is gonna, what it's going to do to our two databases. But one thing is for sure, I believe with quantum sensing, quantum sensors will just have so much more data to deal with. And that will challenge database technology, or Big Data technology in itself, right?

我们很难弄清楚量子科技会对数据库造成什么影响。但有一点是肯定的，我相信随着量子传感器的应用和普及，我们将有更多的数据需要处理。而这将对数据库技术或大数据技术带来挑战。

Then when it comes to go a little bit beyond 20-30-50 years, maybe or 50 years a bit later. But yeah, one of my favorite topics DNA storage basic for the store information to store data within synthetic DNA. And this is interesting, because we know essentially has been talking about numbers initially how much data we have

再过20、30、50年，甚至超过50年之后，就得谈到我最喜欢探讨的话题之一，DNA存储，也就是在合成DNA中存储数据。这个话题很有意思，因为封神刚刚一直在谈论我们目前拥有海量数据。

a lot of this data we don't look at every day, right, we store it in the long from the long term, because we need to for the law says we have to keep records around for hundreds of years, right.

很多数据我们并不是每天查看，只是长期保存而已，因为法律规定我们必须保存数百年的数据记录。

And we do this with traditional technology with tape, disk, they don't last forever, the last maybe 10,15 years. And then we need to copy the data on to a new disk or a new tape etc. So as always, as data migration, as much as the hassling is also quite expensive. And a lot of companies don't want to afford this anymore, can't afford to do this anymore.

我们使用磁带和光盘等传统技术保存数据。但是磁带和光盘无法永久保存，可能顶多保存10到15年。接着，我们就需要把数据复制到新的光盘或磁带上。数据迁移耗时耗力耗财，很多公司要么不想再承担这样的成本，要么承担不起这样的成本。

So what we're looking at with DNA storage, for example, is really to store data for 10s of years, maybe hundreds of years, right, such that we can retrieve it.

通过DNA存储技术，我们希望将数据存储几十年，甚至几百年，以便日后检索。

So we really can take the data, convert it to two strings of nucleotides and then synthesize this and store it in, in the fridge essentially. And when we need it, we sequence and get it back. So anyway, that's kind of I think that's going to happen.

我们可以把数据转换成核苷酸串，然后合成并储存在冰箱里。需要时，我们再进行测序，并取回数据。这就是我所设想的未来。

So generally, I don't want to focus too much on DNA storage itself, I think like, the underlying technology will change drastically 总体来讲，我不想过多地关注DNA存储本身，我认为其底层技术将会发生翻天覆地的变化。

in the past, we looked a lot of when we looked at storage, the storage medium, we had a lot of collaboration with computing, and electrical engineering. Now I think we're getting to a point where we go from, from computing, collaborating between computing and biology, or chemistry, etc.

过去研究存储介质的时候，我们与计算和电气工程领域有很多合作。而现在，我们开始在计算和生物或化学等等领域之间进行协作。

Doesn't have to be DNA can be another kind of storage medium. But I think that's what's going on.

不一定是DNA，也可以是另一种存储介质。但我的设想大概就是这样。

And what's quite interesting there is also I think, when we look at a little bit beyond 20 years, when it comes to DNA storage, but we can also implement some of some data processing some data analytics on top of the DNA using biological processes,

同样有意思的是，展望20年之后，在DNA存储方面，我们还可以通过生物过程在DNA之上实现数据处理和数据分析。

which is extremely energy efficient, and also very, very fast.

这种做法非常节能，而且速度极快。

There are limits to this technology, but we'll find out over the next couple of years. next couple of decades, maybe we'll come in on there

这项技术存在局限性，我们将在未来几年或几十年内找到答案。

but I think generally that we also will the whole field of computing will expand into other into other will collaborate more with other fields. And that also has implications for databases for data analytics to

can use biological processes or chemical processes or anything or similar to do computations right. I think that's that's what's gonna that's what's definitely gonna happen. But it's, you know, the difficulty in the future is very difficult to predict.

总体而言，整个计算领域将会扩展到其它领域，与其它领域开展更多的协作。而这也将对数据库和数据分析产生影响，我们可以利用生物过程、化学过程或其它类似的方法进行计算。我想这绝对是一个趋势。但未来是很难预测的。

The implication of quantum, for example, like I mentioned, quantum sensing will deliver tons of data. But there's gotta be other implications.

例如量子技术的影响。如我之前所言，量子传感技术的应用和普及将给我们提供大量数据。但除此之外，肯定还会产生其它影响。

For example, it's one tiny operation in a database, query optimization, which is kind of like you give the database a query, it figures out how to do it efficiently. And that takes a lot of time to compute to figure out how to execute that query efficiently.

例如，数据库中有这样一个小操作，即查询优化，也就是说，你在数据库里进行一项查询，它会找到高效执行的办法。这一操作需要花费大量的时间。

And we've also already seen in the community that somebody took a query optimization and implemented it on a quantum computer showing that this would be massively faster to optimize the query on the quantum computer. So there's a lot of really I don't think I understand all the implications of quantum but there the quantum computing but that will be definitely also have an impact on databases.

而我们已经社区里目睹了这样一个案例，有人在一台量子计算机上进行查询优化，结果表明，在量子计算机上优化查询，速度要快得多。我无法了解量子技术的所有影响，但量子计算肯定会对数据库产生影响。

So in the short term, adding to essentially I say, is really kind of I think AI is having a tremendous impact in the short term, in the somewhat longer term, I think, we really have to think about interfaces to data virtual reality being one of them, right? Augmented reality being another, but we need to think about how can we make it easy for people to interact with data and understanding typing the query that works for an analyst that's not going to work for everyone right for pretty good for for a broad class of people who need to you know, I think we all need to deal with interpret and analyze data and I think we need to make it easy for everyone. That's a little bit more medium term and in the long term, I think that hardware will change dramatically with quantum with DNA storage with other types of storage medium etc. But 100 years I'm not gonna make a prediction here that's too far out.

短期而言，接着封神刚刚的观点讲，我认为人工智能将在短期内产生巨大影响，而更长远来讲，我们必须思考数据界面，比如虚拟现实和增强现实。我们必须思考如何找到更简单的方法，帮助公众与数据互动，理解数据。输入查询对数据分析师而言是可行的，但并不适用于所有人。所以，我们需要解释和分析数据，降低数据的门槛。这是针对中期而言。长远来看，在量子技术、DNA存储和其它类型的存储介质影响下，硬件将发生巨大的变化。但100年后我就不做预测了，那太遥远了。

主持人：感谢2位朋友，本期节目的最后，我们也为数据库团队和Dr. Heinis打个招聘广告。

Databases research falls within the expertise of Mr. Heinis. Let's invite Mr. Heinis to share with us the prospects of database research in academia. I would like to thank our two friends. I'd like to take this opportunity to share a recruitment ad for the database team and Dr. Heinis.

封神：对数据库技术有热情的，有技术理想，且技术过硬的同学。具体数据库TP、NoSQL、AP各个方向都在招聘。目前我在阿里云数据库重点做数据湖分析，欢迎大家联系我。封神：dragon.caol@alibaba-inc.com

托马斯·海尼斯：Currently, the specific sub-fields of databases, including TP, NoSQL and AP are all recruiting talents. We welcome those who are passionate about database technology, who have technical aspirations, and who are technically proficient. Currently I focus on DLA at Alibaba Cloud. Look forward to hearing from you. Fengshen: dragon.caol@alibaba-inc.com.

Absolutely, of course I do. We have Imperial has been very good collaborations with China in general. And we, for some reason, I don't know why. But we have a big share of Chinese students though, and they are very talented. So if any one of those would like to work on kind of and we offer everything, you know, internships, student shapes for PhD and postdocs as well, if anybody wants to work to to change database technology in the future, of course, you can go to Alibaba or you can come You know, seriously, I'm really looking always recruiting interested to the students, we look at all kinds of aspects of databases, like I mentioned. So, one of the some of the topics that I work on with my team are AI, virtual reality and DNA storage, but we also have other aspects. So, if anybody wants to kind of you know, learn learn these technologies work with these technologies and contribute to this research please do get in touch.

当然了。帝国理工学院和中国一直保持着非常良好的合作。出于某种原因，我也不知道是为什么，我们的学生中有很很大一部分是中国学生，他们才华横溢。假如他们有兴趣参与研究，我校提供各种实践机会，招收实习生、博士和博士后等等。如果你想要改变未来的数据技术，你可以选择加入阿里巴巴集团，或者成为帝国理工学院的一份子。我一直在寻找有志于此的学生。正如我一开始提到的，我们关注数据库的方方面面。我们团队正在研究的课题包括人工智能、虚拟现实和DNA存储等技术，但也包括其它方面。如果有人想学习这些技术、使用这些技术并为这类研究做出贡献，欢迎联系我们。

原文链接：<https://developer.aliyun.com/article/781472?>

**版权声明：**本文内容由阿里云实名注册用户自发贡献，版权归原作者所有，阿里云开发者社区不拥有其著作权，亦不承担相应法律责任。具体规则请查看《阿里云开发者社区用户服务协议》和《阿里云开发者社区知识产权保护指引》。如果您发现本社区中有涉嫌抄袭的内容，填写侵权投诉表单进行举报，一经查实，本社区将立刻删除涉嫌侵权内容。