从Web2到Web3：数据规模带来的新问题

作者：Kerman Kohli 来源：substack 翻译：善欧巴，本站

现在是 2024 年，你会认为获取加密数据很容易，因为有了 Etherscan、Dune 和 Nansen，你可以随时查看想要的数据。表面上确实像这么回事儿。

## 规模

你看，在正常的 web2 领域，当你的公司有 10 名员工和 100,000 名客户时，你产生的数据量可能不超过 100 GB（处于上风）。这个数据规模足够小，你的 iPhone 可以处理你的任何问题并存储所有内容。然而，一旦你有 1,000 名员工和 100,000,000 名客户，你处理的数据量可能现在有数百 TB，甚至 PB。

这从根本上来说是一个完全不同的挑战，因为你要处理的规模需要更多的考虑。要处理数百 TB 的数据，你需要一个分布式计算机集群来发送作业。在发送这些作业时，你必须考虑：

如果工人未能履行职责会发生什么情况

如果一名工人比其他工人花的时间长得多，会发生什么情况

你如何确定给哪个工人分配哪个工作

如何将所有结果合并在一起并确保计算正确

这些都是在处理跨多台机器的大数据计算时需要考虑的事项。规模会产生一些问题，而这些问题对于那些不使用它的人来说是看不见的。数据是这样的领域之一，规模越大，需要的基础设施就越多，才能正确管理它。对大多数人来说，这些问题是看不见的。要处理这种规模，您还面临着其他挑战：

知道如何操作这种规模的机器的极其专业的人才

存储和计算所有数据的成本

前瞻性规划和架构，以确保您的需求能够得到支持

有趣的是，在 web2 中，每个人都希望数据公开。在 web3 中，终于可以公开了，但很少有人知道如何做必要的工作来理解它。一个欺骗性的事实是，通过一些帮助，你可以相当轻松地从全局数据集中获取你的数据集，这意味着“本地”数据很容易，但“全局”数据很难获得（与每个人和每件事有关的东西）。

## 碎片化

好像事情已经不具有挑战性了，因为你必须处理的规模。现在有一个新的维度让加密数据变得具有挑战性，那就是由于市场的经济激励，加密数据会不断分裂。例如：

新区块链的兴起。目前有近 50 个 L2 已上线，50 个已知即将上线，还有数百个正在筹备中。每个 L2 实际上都是一个需要索引和配置的新数据库源。希望它们是标准化的，但您不能总是确定！

新型虚拟机的兴起。EVM 只是一个领域。SVM、Move VM 和无数其他虚拟机正在进入市场。每种新型虚拟机都意味着一种全新的数据方案，必须从基本原理和深入理解的角度来考虑。有多少虚拟机？投资者将以数十亿美元的投资来激励新事物！

新账户原语的兴起。智能合约钱包、托管钱包、账户抽象化为您实际解释数据的方式带来了新的复杂性。发件人地址可能实际上不是真实用户，因为它是由中继提交的，而真实用户可能在混合中的某个地方（如果您仔细查看的话）。

由于你无法量化你不知道的东西，碎片化可能特别具有挑战性。你永远不会知道世界上存在的所有 L2 以及总共会出现的虚拟机。一旦它们达到足够的规模，你就能跟上，但这是另一个故事了。

## 开放，但不可互操作

我认为最后一个问题让很多人感到惊讶，那就是数据是开放的，但不能轻易实现互操作。你看，团队拼凑起来的所有智能合约就像一个大型数据库中的小型数据库。我喜欢把它们看作模式。所有的数据都在那里，但开发智能合约的团队通常知道如何将它们拼凑在一起。如果你愿意，你可以花时间自己去理解它，但你必须对所有潜在的模式进行数百次这样的操作——而且，在没有交易另一方买家的情况下，你怎么能不花大笔钱就做到这一点呢？

如果觉得这个太抽象了，我来举个例子。你说“这个用户使用桥梁的频率是多少？”。虽然这看起来是一个问题，但里面嵌套了很多问题。让我们来分解一下：

首先，你需要知道所有存在的桥梁。还有你关心的链。如果是所有链，那么我们上面已经提到了为什么这很有挑战性。

然后，对于每座桥梁，你需要了解它们的智能合约是如何运作的

一旦你理解了所有的排列，你现在需要通过一个可以统一所有这些单独模式的模型来推理

上述每个挑战都很难解决，而且需要大量的资源。

## 结果

那么这一切会导致什么呢？好吧，我们今天的生态系统状况是……

生态系统中没有人真正知道真正发生了什么。只有难以正确量化的活动概念。

用户数量虚高，女巫攻击难以检测。指标开始变得无关紧要和不可信！真假对市场参与者来说甚至不重要，因为它们看起来都一样。

使链上身份真实化的主要问题。如果你想拥有强烈的身份意识，准确的数据至关重要，否则你的身份就会被歪曲！

我希望本文能帮助您了解加密数据领域的现实情况。