



GenieAnalytics

解碼5G大數據：全網流量分析實務應用與安全防護新視野

Jacob Chiang
CTO, Genie Networks



“We wanted flying cars, instead we got 140 characters.”

Peter Thiel, Zero to One: Notes on Startups, or How to Build the Future

- 網際網路的普及
 - 資訊獲取
 - 強化溝通
 - 電子商務與線上服務
 - 遠端工作與教育
 - 娛樂與休閒
 - 自動化與物聯網
- 網路犯罪的威脅
 - 網路詐騙
 - 網路釣魚
 - 惡意軟體
 - 網路霸凌和騷擾
 - 資料外洩

- 分析網路封包內容，而不僅是標頭資訊。
- 識別應用程式流量。
- 識別異常行為。
- 及時發現並阻止潛在攻擊。
- 提供詳細流量紀錄，助資安人員了解網路使用情況。
- 提供有力證據，協助執法單位打擊網路犯罪。

- 電信全網流量

- 每秒約800萬條流量紀錄
- 每年約200兆條流量紀錄

- 比較數據

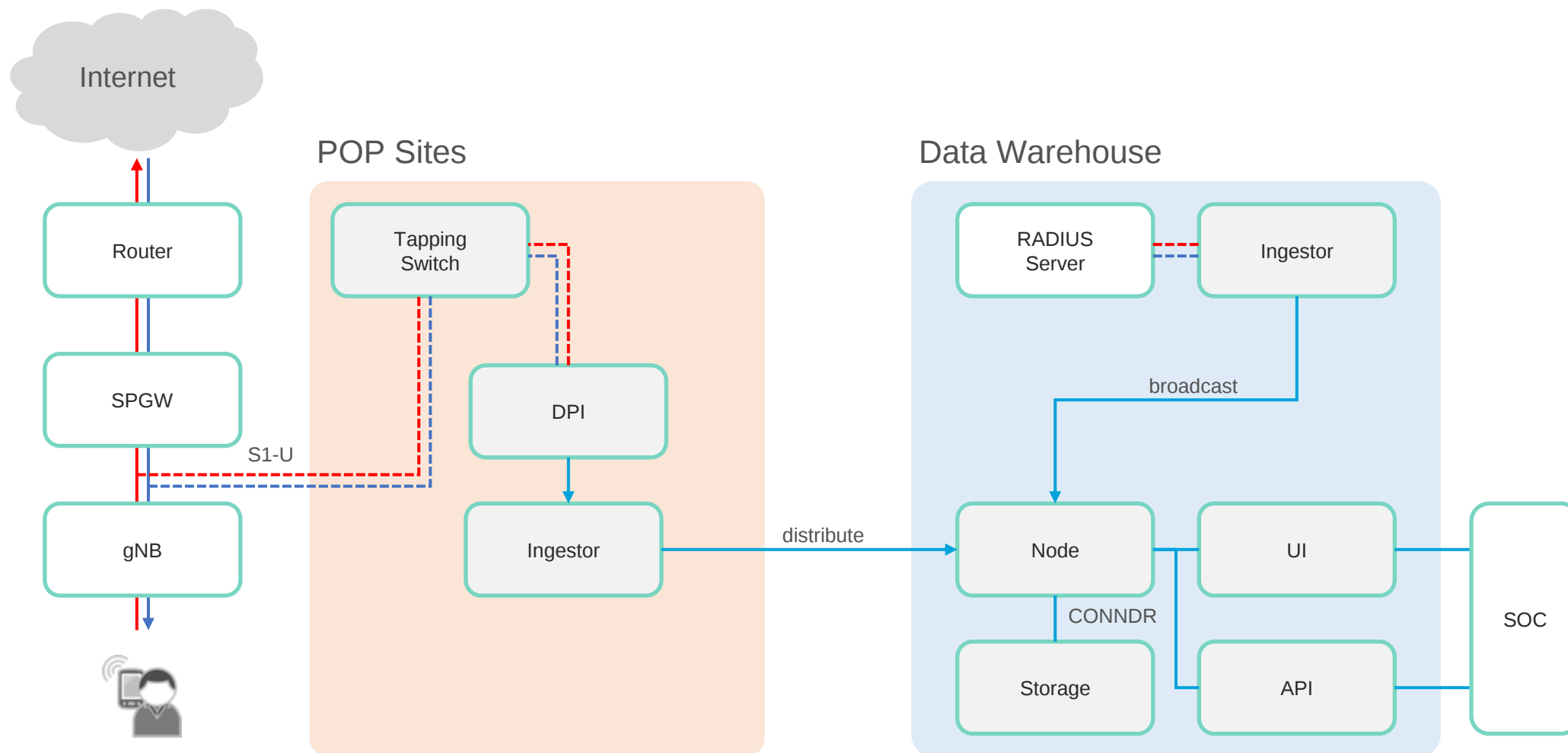
- 台灣每年約80億張發票
- Google每年約6兆次搜尋

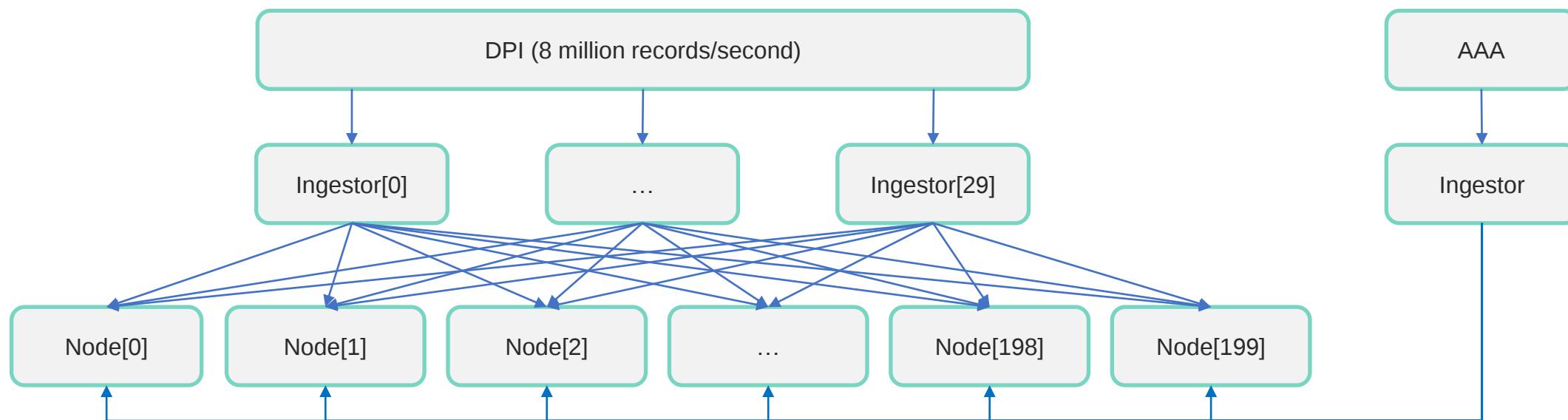
- 技術挑戰

- 異質資料關聯
- 海量資料儲存
- 長期資料保存
- 高速資料查詢

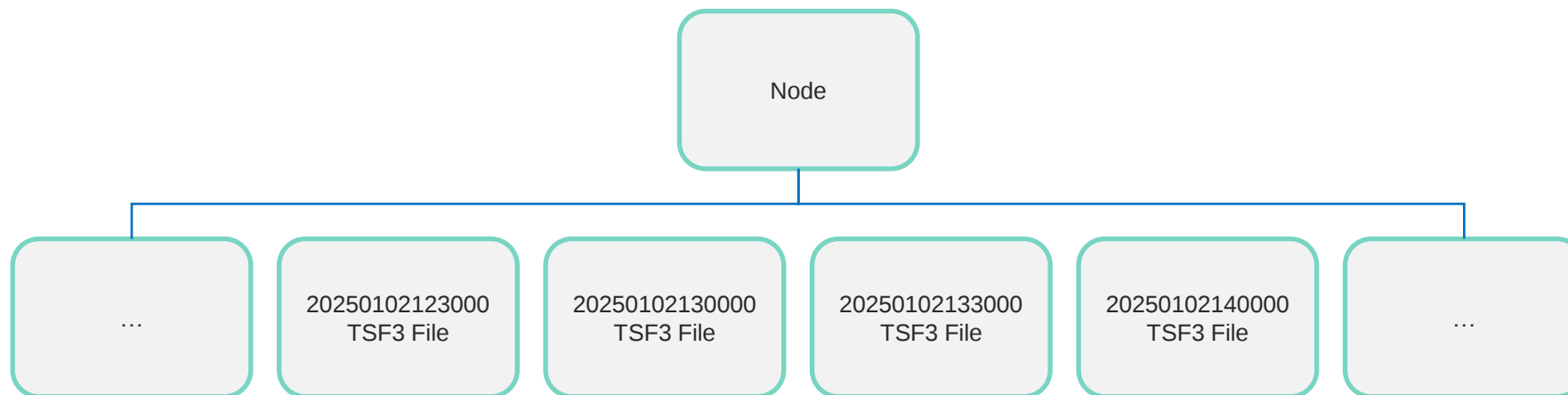
全網流量紀錄系統架構

GenieAnalytics





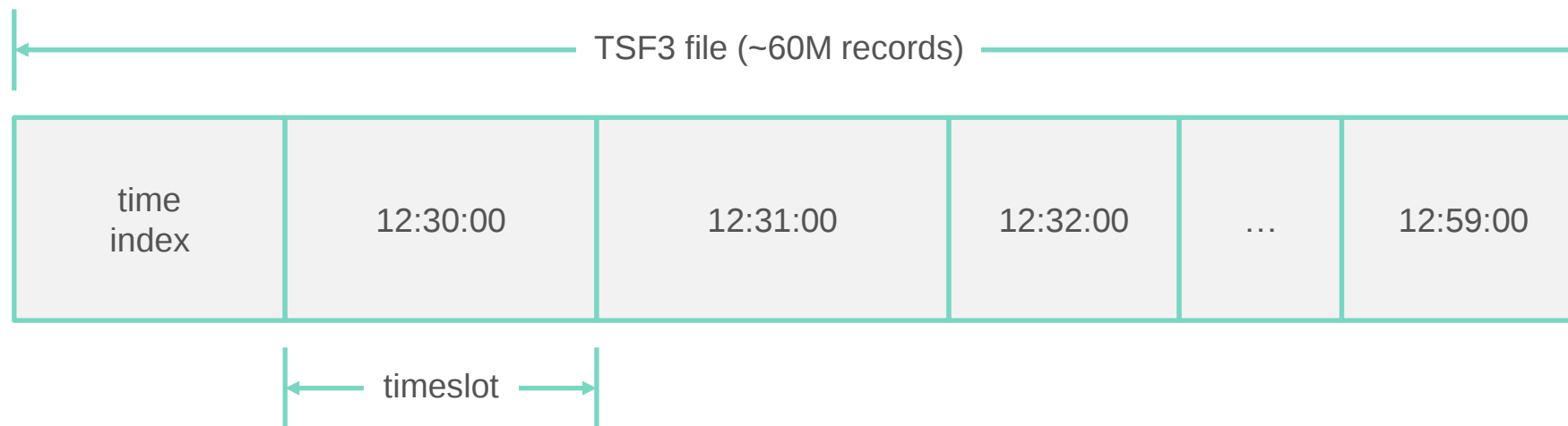
- 將200兆筆流量紀錄平均分配至200台資料節點
- 每個資料節點儲存約1兆筆流量紀錄(約200TB)



- 每30分鐘的流量紀錄(約6000萬筆)儲存在一個資料檔
- 定期刪除過期資料檔而非流量紀錄，大幅簡化保存設計

高速資料查詢 – 時間索引

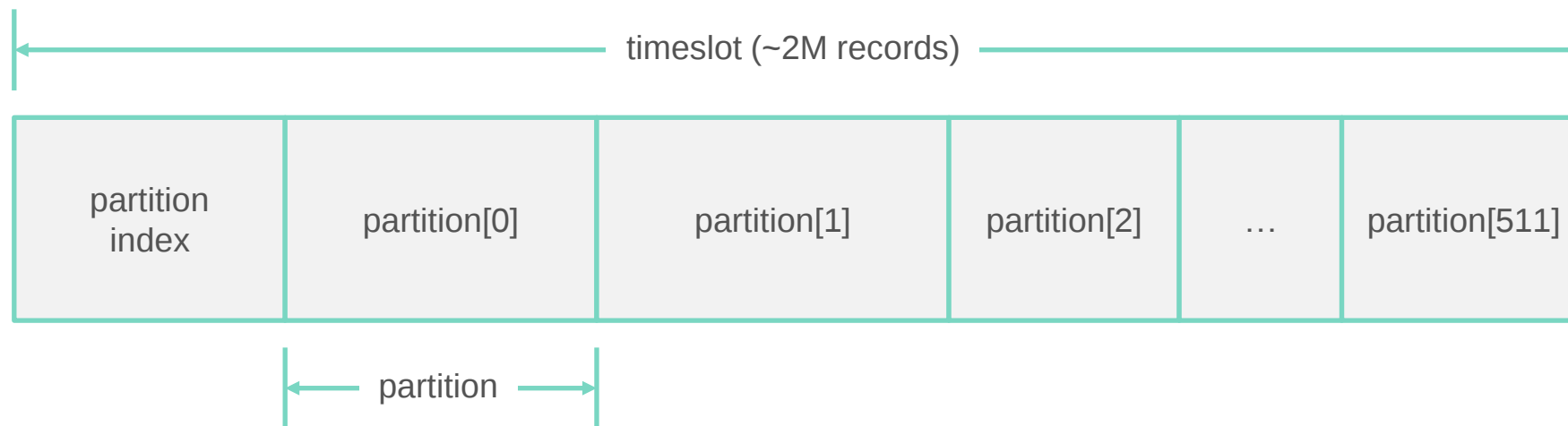
GenieAnalytics



- 同一分鐘的資料放在同一時間槽(約200萬筆)
- 頭部提供時間槽索引以快速定位查詢時間區間

高速資料查詢 – 分割索引

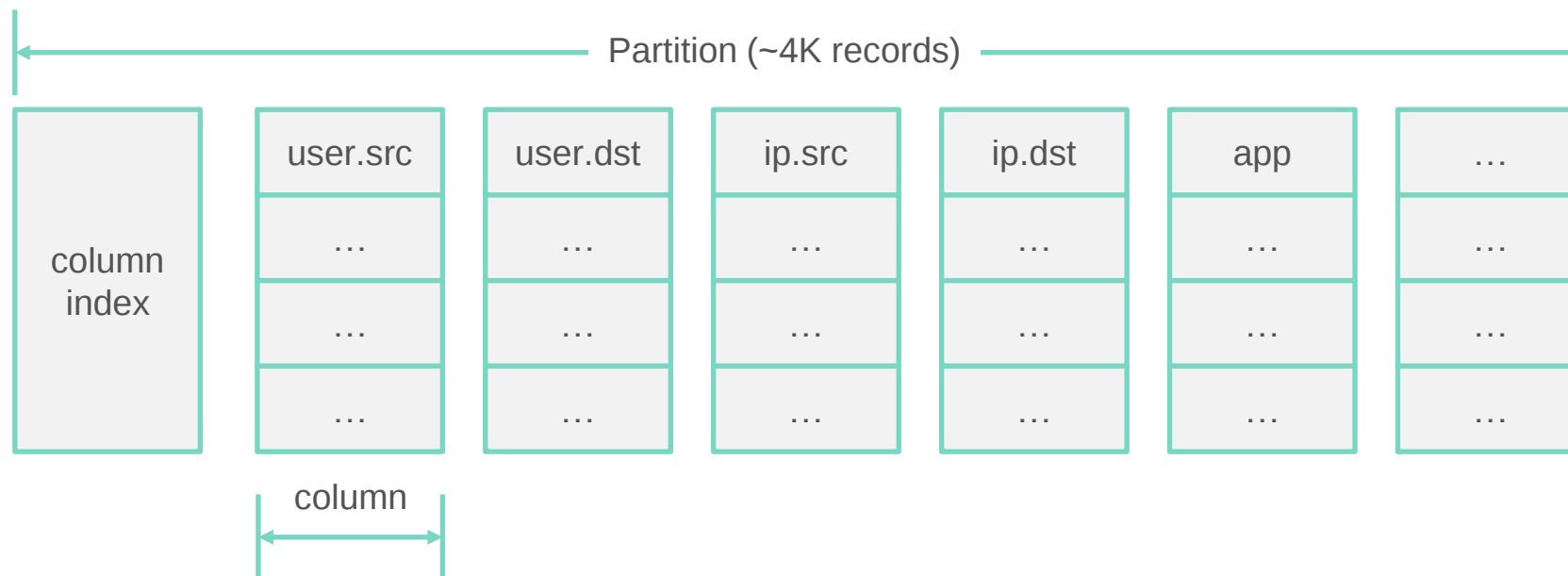
GenieAnalytics



- 將門號分配至512分割，每分割約4000筆資料
- 頭部提供分割索引以快速定位門號所屬分割

高速資料查詢 – 分欄儲存

GenieAnalytics

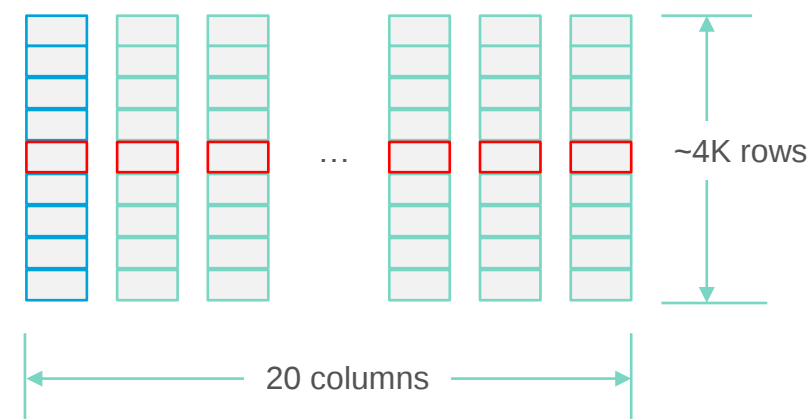


- 將約4000筆資料分欄進行壓縮，然後寫入硬碟
- 頭部提供分欄索引以快速定位比對欄位

高速資料查詢 – 分欄加速

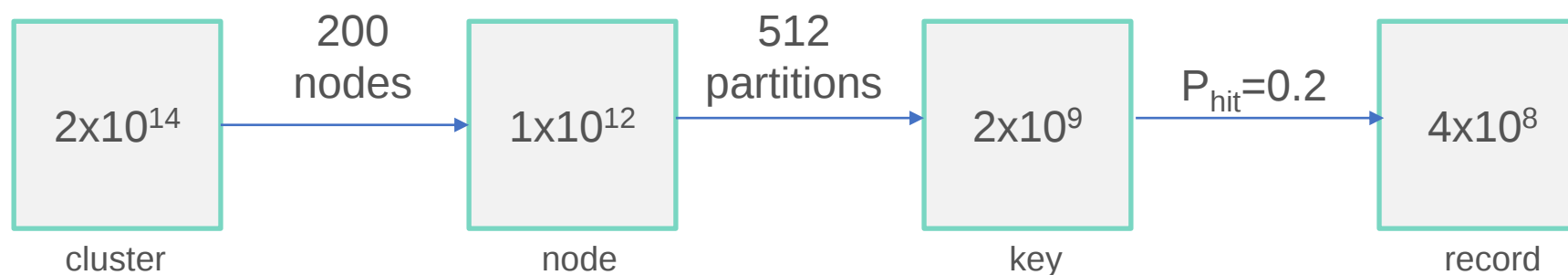
GenieAnalytics

- 每一區塊有4000筆資料，每筆有20欄
 - 掃描時只讀取比對欄
 - 比對欄命中才讀取其他查詢欄
- 假設查詢1000萬用戶其中一用戶資料
 - $P_{hit} = 1 - (1 - \frac{1}{10000000})^{4000} = 0.0004$
 - 幾乎所有區塊都只讀取比對欄
- 假設查詢Class B地址段其中一IP資料
 - $P_{hit} = 1 - (1 - \frac{1}{65536})^{4000} = 0.06$
 - 僅有6%區塊要讀取其他查詢欄



高速資料查詢 – 順向查詢

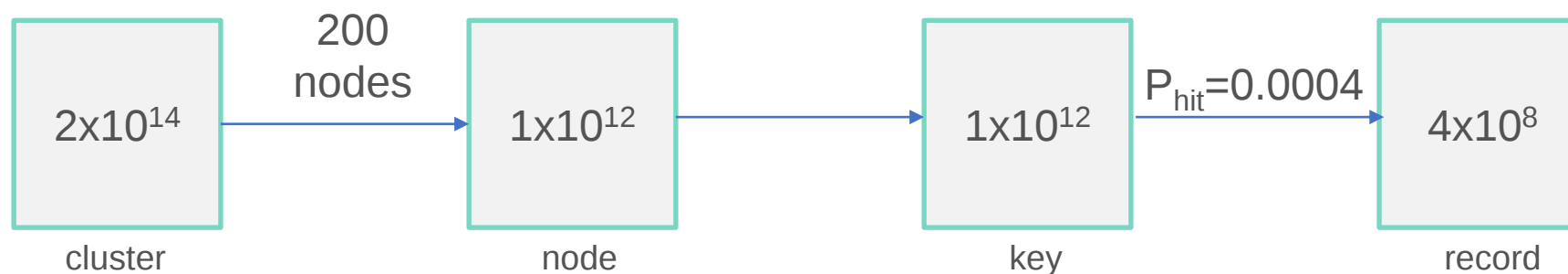
GenieAnalytics



- 200台主機同時查詢，每台主機查詢約1兆筆資料
- 門號分配至512分割，每台主機實際比對約20億筆索引
- 區塊命中率約20%，每台主機實際讀取約4億筆紀錄

高速資料查詢 – 反求查詢

GenieAnalytics



- 200台主機同時查詢，每台主機查詢約1兆筆資料
- 無法用門號加速，每台主機仍需比對約1兆筆索引
- 區塊命中率約0.04%，每台主機實際讀取約4億筆紀錄

- 全網流量紀錄的數據規模遠超想像
- 資料倉儲的關鍵在不斷進行「降維」
 - 使用資料叢集技術將關聯、儲存、查詢分散至多台主機
 - 將不同時間資料分散至不同資料檔，可大幅簡化長期保存的設計
 - 用時間作為一級索引，跳過非查詢區間以減少讀取資料
 - 用門號作為二級索引，跳過非查詢門號以減少讀取資料
 - 用欄位作為三級索引，跳過非查詢欄位以減少讀取資料

THANK YOU!

• www.genie-networks.com

