Graph Analytics on PB-Scale Datasets: Future AI

Scenario:

Dr. Smith, a renowned oncology specialist, is about to meet with Mr. Brown, her new referral patient. *"How are you Mr. Brown?"* She opens up the conversation as she enters the room. As Mr. Brown shares his concerns, Dr. Smith deliberately listens, entering notes in her tablet. A Clinical Decision Support (CDS) app on the tablet recommends her three care paths (treatment courses) for the patient. However, with each option, there are additional data points needed and the screen flashes with the message, "More data needed".

Dr. Smith asks questions about past drug interactions, allergies, extended family history, past-7 day symptoms, and other information that's not readily available in his health record. She enters the data in the follow-up screens, and answers generate new follow-up questions. Finally, as all the data is entered, CDS provides a recommendation for a care path **A** with optimal outcome. Dr. Smith informs Mr. Brown that after comparing his clinical data with thousands of similar patients, that using Care Path **A** predicted the highest probability of positive outcomes. Dr. Smith chooses it and directs Mr. Brown to the genomic testing blood draw, so she can use this information in adjusting the dosages of drugs depending on Mr. Brown's expected metabolic rate of absorption.

Is this a Farfetched scenario? Hardly.

This is a *simple version* of a scenario that Dan McCreary, DE of United Health Group's (Fortune #7) Advanced Technology Collaborative is looking to employ to drive best possible patient outcomes at the least cost with empathetic and engaging relationship between the doctors and patients.

Role of Graph Analytics in the Use Case:

As Dr. Smith deliberately listens, entering notes in her tablet (activating speech-to-text entry), her patient data is populated in the internal **Peta-Byte size** datastore which consists of hundreds of millions of health records. Data is stored as a Graph because a relational database cannot handle the data at scale *and* meet the real-time query response time of 200 ms. This is because as the relational data scales, new schemas need to be created, queries become multi-nested and the latency increases exponentially – Graphs don't suffer from this issue due to ease in adding new nodes on-the-fly and inherent relationships embedded as part of the data. The CDS App **traverses the Graph** using first the **Random Walk** and **Louvain algorithm** to identify clusters of similar patients, then **Cosine Similarity** to identify similar patterns of symptoms, duration, age, gender, and so forth to recommend viable care paths. The app indicates where more data is required, creating a real-time patient engagement to identify the best possible care path. With additional data, the peta-byte scale Graph is populated and queried on-the-fly to find further matching patterns with optimal outcomes for the patient.

This diagnosis is dynamic – the traversal path depends on the answer to previous questions – and needs to have a response time of **less than 200 ms** to make the doctor-patient conversation feel natural. New questions that emerge must be answered before the Graph can be traversed further. This is called **data dependent pointer hopping**. Analytics finds the pattern that matches patient conditions, while the real-time Q&A drives recommendations for the care path that works best for this patient's personal medical history, genotype, and other relevant data. This recommendation and treatment outcome in turn become part of the total Graph data that can be used in diagnosing future patients. Further, Dr. Smith uses the pattern search based on **Random Walk, Louvain** and **Cosine Similarity** to determine which drug and how much of it should be given to Mr. Brown depending on his DNA profile.

The above scenario is one of numerous scenarios driving graph analytics in healthcare. Numerous other use cases are seen in other verticals such as Cybersecurity (~6 PB today), Drug Discovery (multi-PB before 2025), and Financial Investments & Fraud Detection (>1 PB today).

What is a Graph?

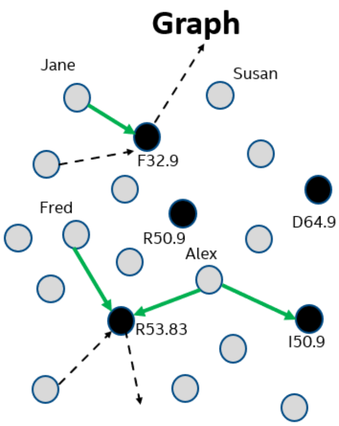
A graph represents data in the context of their information relationships as opposed to a relational database which represents information as schemas. All the individual events in data is represented as nodes (vertices), while the relationships among events is described via links (edges) between nodes. As the new data is added, these relationships and nodes can easily be created and manipulated **without creating a new schema** – a fundamental limit to RDBM scalability for these highly variable and irregular relationship properties. In addition, the relationships among data elements can be easily and much more rapidly **(in milliseconds)** gleaned via traversing the graph rather than doing expensive high latency **(in minutes, hours, days)** SQL-based SELECT and JOIN operations across multiple tables in relational databases at this scale. The Figure below shows an example of data in relational database and its equivalent Graph representation.

Relational Database

ICD-10 Diagnosis Codes (100K of them...)

	D64.9	F32.9	R53.83	I50.9	R50.9
Jane Doe	0	1	0	0	0
Fred Smith	0	0	1	0	0
Susan Brown	0	0	0	0	1
Alex Johnson	0	0	1	1	0

99% of the values are zero



Why Should Intel/DPG Care About Graph Applications?

There are signals that point to the opportunity.

- Motleyfool.com: Cybersecurity leader CrowdStrike believes cloud-security spend will jump from **$1.2 billion in 2020 to $12.4 billion in 2023** -- a tenfold increase. They are differentiating with Graph Analytics collecting and analyzing over 4 trillion data points per week. This enormous Peta-Byte scale database, combined with sophisticated AI-powered analytics, allows CrowdStrike to be well positioned to capture market share.
- ZDnet: "Nvidia's cuGraph's bet is to make graph analysis ubiquitous. Doing this would not only mean **faster analytics but potentially a stepping-stone in the future of AI**, which, to a large extent, goes through graph."
- Andy Jassy, AWS: "Graph databases are finding increased adoption on the cloud and **traditional over-reliance on relational databases is slowly breaking down.**"
- Researchandmarkets.com: Graph Analytics Market is expected to reach **$2.03 billion by 2027, growing at a CAGR of 19.1%** during the forecast period of 2019 to 2027. The ability to **analyze low-latency queries** and the ability of graph analytics to **uncover relationships between data in real-time** are some of the key factors driving the growth.

Why Can’t Xeon Address This Market?

The **sources of advantages of Xeon** for regular applications become **sources of disadvantages** for graph. The differences between current applications and graph applications pose challenges in that Graphs are primarily data-dependent pointer hopping applications whose performance is dependent on traversed edges/sec (TEPS) in a dependent-pointer chasing pattern, as opposed to peak dense FLOPS. Current architectures are FLOPS optimized.

1. Graph applications operate at Peta-byte scale. The dataset must be distributed though many nodes, and data could be anywhere in this distributed memory. The overhead of network accesses in the current architectures is a huge penalty to transfer the memory to the compute node.
2. Graph Applications do not lend well to cache and pre-fetch since the data is nearly random in distribution across the networked machine nodes. The current memory architectures are optimized for locality of data.

Why Did DARPA Fund PIUMA?

DARPA started Intel’s PIUMA program with 50+ Graph algorithms with the goal of real-time insights from Peta-Byte scale data with 1,000x performance improvement over traditional CPU and GPUs. DARPA also sought a significant Perf/watt breakthrough as well. Program goal was to “provide the rocket for the advancement of graph analytics to open new pathways for understanding an ever-increasing torrent of data, allowing analysts to draw conclusions from the patterns in the data which previously there would have been no hope of answering”.

Random Walk, Louvain and Cosine-Similarity are part of the DARPA algorithms. Our preliminary analysis shows, as an example, a Random Walk algorithm on 1 PIUMA node is 10x faster than a DGX-1 GPU system and 16 Xeon sockets. Scaling further, executing Random walk on 16 PIUMA nodes is 154x faster than a DGX-1 GPU system.