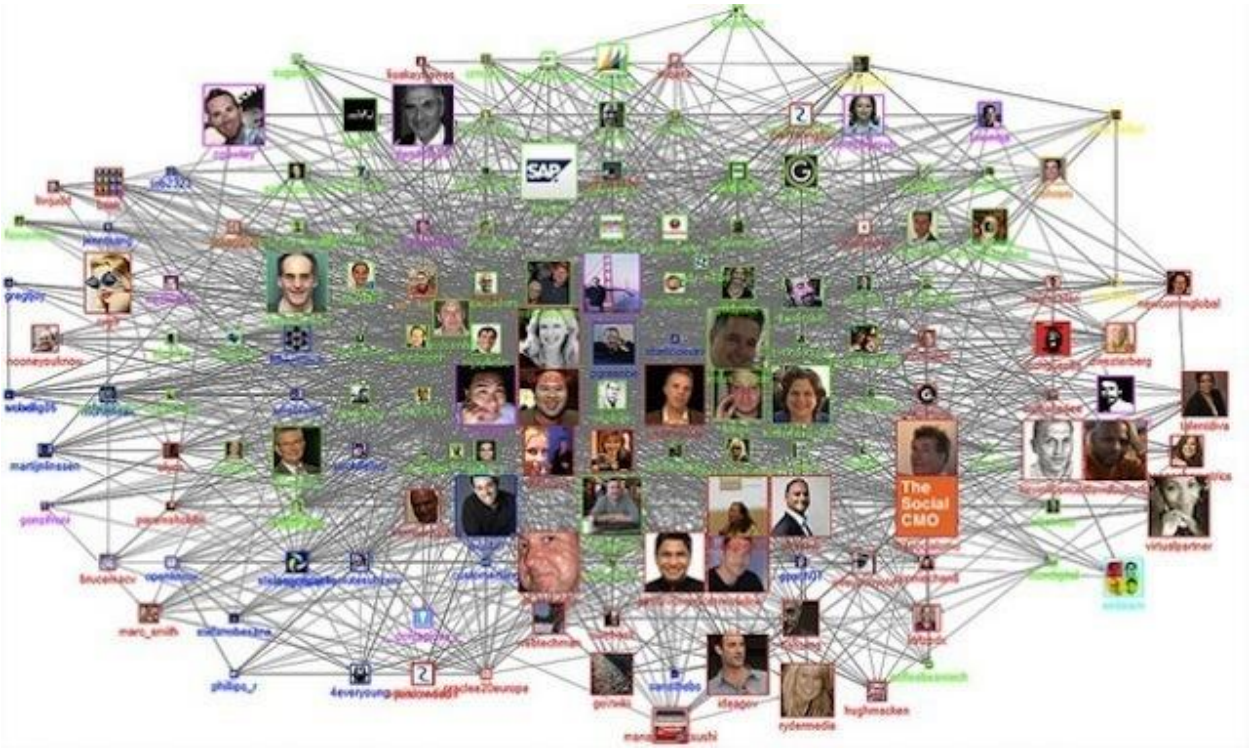
Memory Driven Compute combined with Graph technology delivers unparalleled Search and Analytics



Introduction

For decades, we have known the value and potential to run applications at memory speeds, only recently have memory densities, interconnects, processing power and prices aligned to deliver Large Memory fabrics (in multi-TB range) on a single system image to deliver in-memory database performance at TB scale and at commercially attainable pricing. As this new capability has become accessible to mass market, it has opened the door to Memory-Driven Computing (MDC) enabling a rapid growth for foundational In-Memory Analytics capabilities. Consider the variety of Workloads previously thought impossible to run in memory, and now possible.

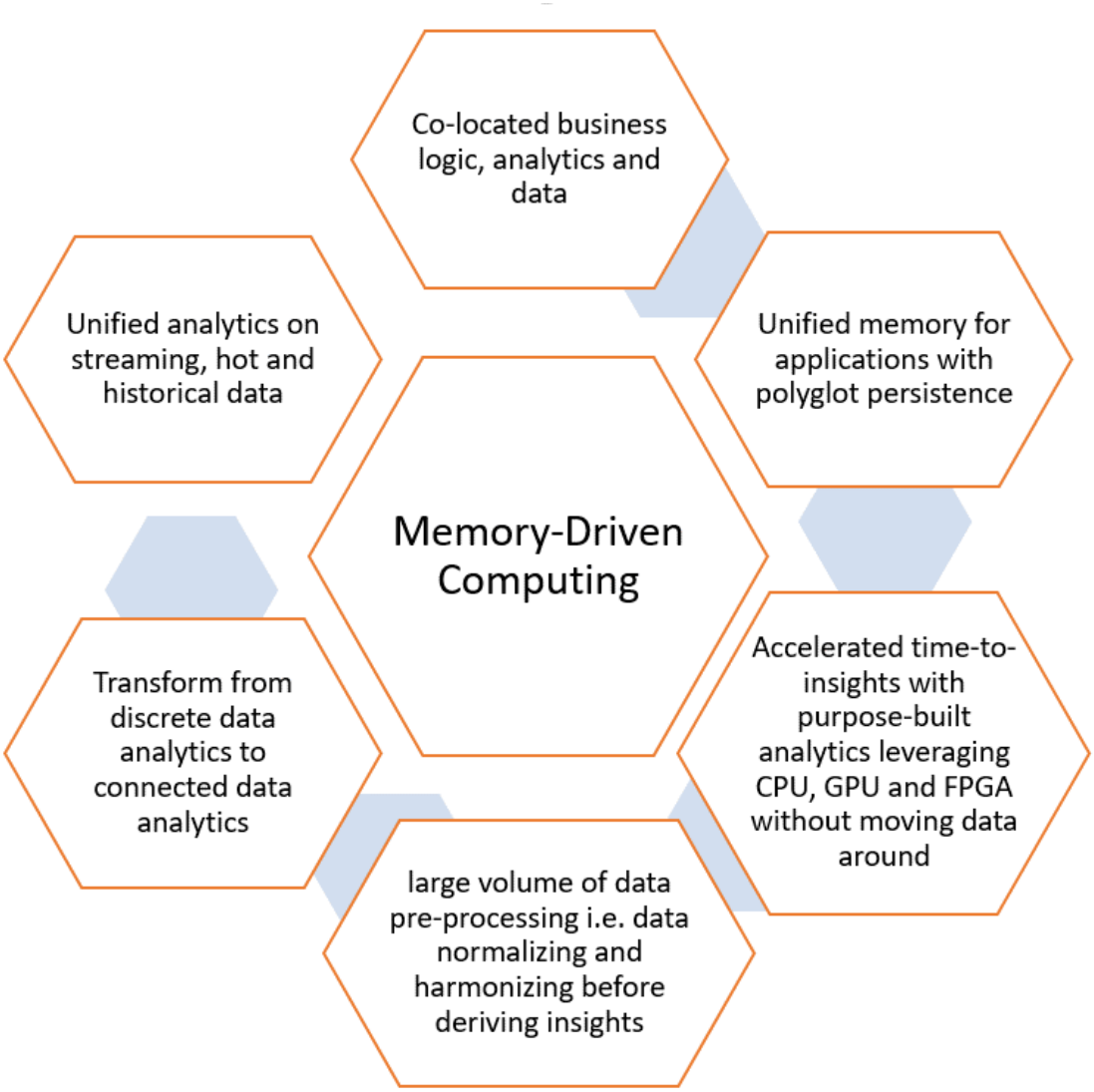
Consider what additional common use-cases can take advantage of In-Memory Analytics with Memory-Driven Computing architecture of large scale up systems like HPE Superdome Flex? Read on to get insights into common use-case and analytics solution with HPE Superdome Flex.

Current enterprise challenges with handling growing volume of data

Businesses are experiencing an exponential increase in data, coming from an explosion of sources and we have a vanishingly small time to turn that data into meaningful action. In past, the fact that data will double every two years led to a very famous law of electronics i.e. Moore's law which stated that transistors will double in integrated circuit every 2 years. This prediction is already proving wrong as data growth and data analytics

requirements are outpacing the compute and storage technologies that provided the foundation of processor-driven architectures for the last five decades.

Deriving such time critical insights, requires architectural shift from compute-driven analytics to memory-driven analytics. This architectural redesign is driven by following challenges and experiments in recent times.



Memory-Driven Computing is an almost infinitely flexible and scalable architecture that can complete any computing task much faster, using much less energy, than conventional systems. The performance of Memory-Driven Computing is possible because now any combination of computing elements can be composed and can communicate at the fastest possible speed – the speed of memory.

Only through a new architecture like Memory-Driven Computing will it be possible to simultaneously work with every digital health record of every person on earth, every trip of Google's autonomous vehicles and every data set from space exploration all at the same time—getting to answers and uncovering new opportunities at unprecedented speeds.

I am going to highlight key New-Gen Analytics workloads taking advantage of Memory-Driven Computing architecture and scaling for real-time analytics requirements.

Next-Gen In-Memory Analytics Solution leveraging Memory-Driven Computing Infrastructure
Let us look at few industry use-cases that take advantage of architectural principals of compute-driven analytics.

Use-Case-1: Cybersecurity Analytics: Cybersecurity Analytics involves identifying cyber intrusion behaviors in a deployed infrastructure comprising of complex network of servers, routers, gateways, storage etc.

Developing such Cybersecurity Analytics involves analyzing massive volume of Infrastructure network traffic information from network connection logs, http logs, dhcp logs, smtp logs, netflow information etc. and there by establish a network of infrastructure entities and relationships. This is achieved by building a network graph which enables detection of network anomaly patterns leading to identifying threats like Zombie reboot, RDP hacking etc.

Typical size of these network graphs comprises of **billions of graph nodes and properties and relationship between graph nodes**. Deriving anomaly patterns across these billions of nodes in real-time requires existence of entire graph in-memory with TBs of large memory infrastructure.

Cybersecurity Analytics use-case with Graph Databases is ideal for memory-driven analytics.

Use-Case-2: Real-Time Recommendation Engine: In this day and age, the need to build scalable Real-Time Recommendations is increasing day by day. With more internet users using e-commerce sites for their purchases, these e-commerce sites have realized the potential of understanding the patterns of the users' purchase behavior to improve their business, and to serve their customers on a very personalized level. To build a system which caters to a huge user base and generates recommendations in real time, we need a modern, fast scalable system.

Deploying a Real-Time Recommendation engine involves In-Memory data processing unifying user data from social-media data sources, existing customer management solutions, existing warehouse historical data etc.

Building such unified analytics platform can be achieved with a combination of Spark based In-Memory data processing and transformed representation of customer data in graph models. These unified analytics built with Spark involves application of massive transformation and action. These phases of transformation and action undergo massive data shuffle operation to pre-process which is very costly in cluster operation.

Typical scale of shuffle spans across TBs of data undergoing Repartition, GroupBy and Join operations which are very expensive owing to frequent I/O operations. Making these TBs of data available In-Memory with large memory infrastructure enables Spark to perform data pre-processing operations faster to meet the requirements of **Fast Data Analytics.**

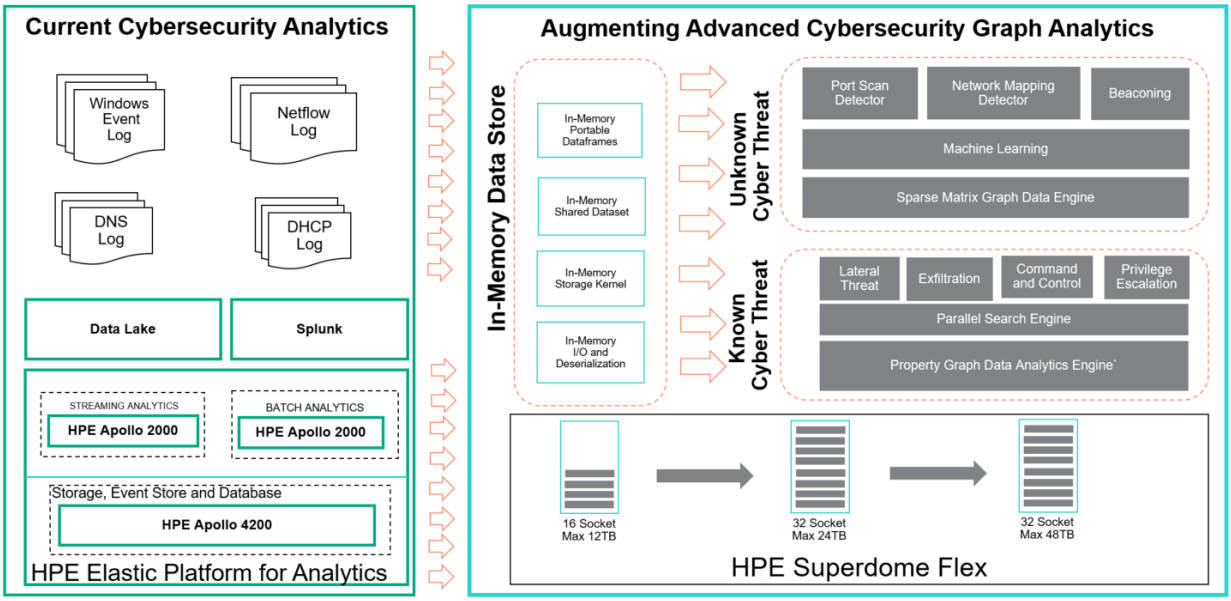
Real-Time Recommendation Engine with Fast Data Analytics is yet another ideal use-case for memory-driven analytics.

Use-Case-3: Fraud Detection: Fraud management has been known to be a very painful problem for banking and finance firms. Card-related frauds have proven to be especially difficult for firms to combat. Technologies such as chip and PIN are available and are already used by most credit card system vendors, such as Visa and MasterCard. However, the available technology is unable to curtail 100% of credit card fraud.

Building a fraud prevention solution requires analyzing credit card transaction in sub-millisecond time frame, detecting outliers in which data-set is verified to identify potential anomalies in the data. With the rise of **machine learning** (**ML**), **artificial intelligence** (**AI**), and **deep learning**, it becomes feasible to analyze massive volume of transactions feeding into enterprise credit card network. These machine learning models are first trained against historical transactions and live inference is achieved by building machine learning pipeline for data acquisition, feature engineering and model serving.

Achieving Real-Time fraud detection against streaming credit card transactions requires building machine learning pipeline in-memory to perform data collection and transformation, feature engineering, hyper-parameter optimization and model serving to make high precision predictions. This can be achieved by implementing Spark and Tensorflow with TensorFrames and taking advantage of co-located GPUs for faster model training.

Building upon these use-cases, below is the high-level architecture for Next-Gen In-Memory Analytics Solution designed to take advantage of scale-up Memory-Driven Computing Infrastructure with HPE Superdome Flex.



<u>Let us talk on some key features of In-Memory Graph Analytics solutions:</u>

1. Adopt and build **<u>large scale In-Memory data ingest</u>** with load capabilities to perform parallel read operation leveraging large memory and multi-processor architecture
2. **<u>Accelerate large scale in-memory analytics with graph query engine</u>** against massive volume of graph data model leveraging large memory and multi-processor architecture
3. **<u>Accelerate connected data analytics leveraging Graph Databases</u>** using search algorithms like Page Rank or Single Source Shortest Path (SSSP) taking advantage of large memory Superdome Flex infrastructure.

Let us look at how In-Memory analytics is achieved leveraging Memory-Driven Computing for real customer implementations.

A large Financial Services organization implemented Cybersecurity Analytics to identify cyber threats in their operational network by detecting network hacks, zombie reboots, RDP hacking etc. One of those workloads

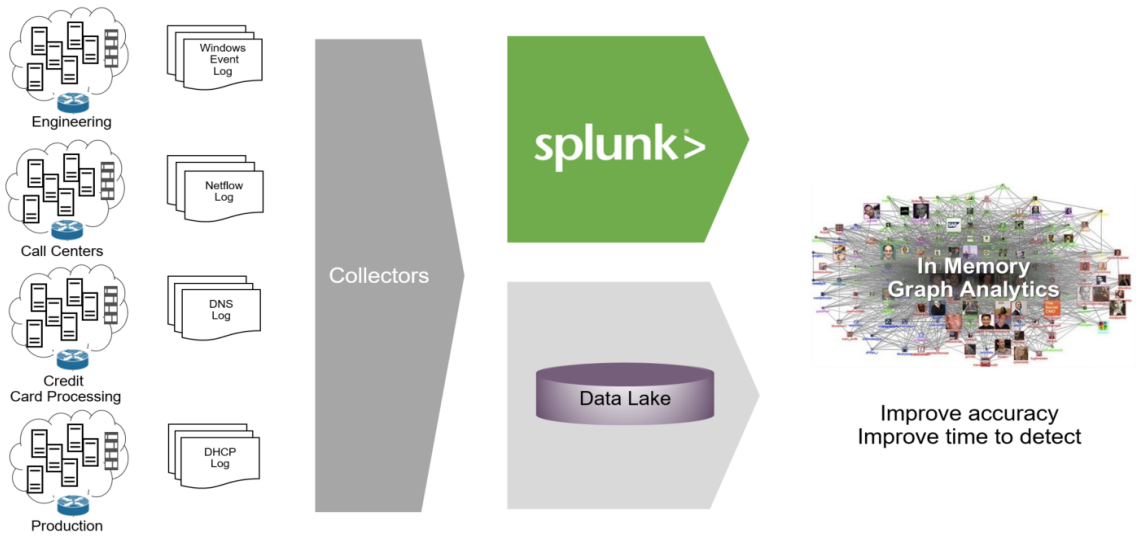
requires a Database. Imagine running a traditional columnar Relational Database in memory, and consider the speed at which you can perform table joins, indexing, search, etc. Walking through memory is 100x faster when the entire table is in memory vs waiting on data retrieval from slower media (HDD or SSD), traversing an I/O Bus, copying into memory and updating pages/pointers. In a traditional RDBMS, while your data is structured, it is not necessarily organized for optimal access, which means multiple data retrievals to walk through to all the data you needed to complete a query request.

Ironically RDBMS should establish and help find Relationships, which they do through the query language and tools and yet the relationships have not really been established until post processing (joins/indexing, or the query). What if we were to look at your data from a different perspective, consider relationships that exist within your data and mapping out those relationships as part of the ingest and database creation.

Graph Databases, on the other hand take this from a different approach, and generate the relationships as data is read in. Graph databases, establish relationships as part of the database creation. Now couple the creation of the database, and in memory database, unleashing a powerful combination. With the database in Memory, and data relationships mapped out in a Graph, walking thru data in memory coupled with relationship based search is 1,000x faster.

To map out the full landscape to attempt to address Cybersecurity threats, multiple data sources need to be combined such as known infrastructure map, system logs, access accts/profiles, network logs. It could take years to map out all the relationships, or design an approach to evolve the relationship functions within a traditional RDBMS. Whereas a Graph Database can dynamically create relationship.

Take as an example, combining various disparate data sources to establish a holistic picture of a company's technology assets and electronic activities. Once combined, searching for known threat signatures is easy. ML and AI techniques can be applied to this combined data, to search for anomalies (unknown and yet to be identified signatures).



Typical landscape of Cybersecurity Analytics would include following key components

a) Host and Network data from deployed network elements is captured in various logs i.e. netflow log, conn log, dhcp log, http logs etc.
b) Critical data about deployed infrastructure and app would also be available in enterprise end-points like DIR Svcs, Usage Data, CMS etc
c) Data from these multiple sources would be collected for data aggregation and discovery as pre-processing for Cybersecurity Analytics
d) Log data is processed to identify Cyber Network elements and there-by build cyber network graph representing Network Elements
e) Data Scientists and Analytics would interact with In-Memory Cyber Graphs leveraging HPE BlueData ML Ops
f) Cyber threat patterns would be identified and analyzed to detect cyber intrusion behavior
g) Develop pre-trained models in the form of Graph Queries to subsequently leverage for Day-Zero Threat Detection

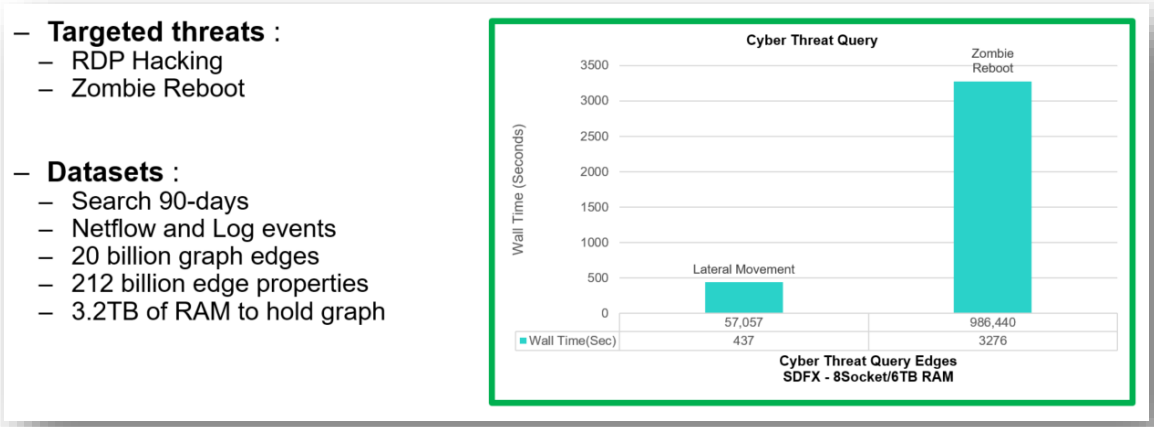
There are multiple challenges required to be addressed while implementing above deployment.

First challenge was to integrate all this data. This was achieved by aggregating all the data from these multiple sources into large memory infrastructure hosting terabytes of data.

Subsequently, aggregated data was transformed into a Graph Data Model and a network graph was built to represent these network entities. Size of network graph included 20 Billion Graph edges (17.9 Billion Netflow Edges and 1.5 Billion log edges) and 212 billion graph edge properties against 3TB of input data from network.

In order to build such network graph, an enterprise software product was implemented in HPE Superdome Flex large memory environment to read the input data from multiple data sources as highlighted above and the tool generated a powerful network graph with billions of edges and edge properties.

Finally, pattern matching operation was performed to detect anomalies. This was achieved by developing complex pattern matching queries. Query response time was measured with deployed in-memory graph.



Achieving such scale of performance for complex cyber threat pattern matching operations, Superdome Flex is well poised enable 6,000+ complex cyber threat detections in one day as against ~6 threat detections in one day in scale-out deployments and there-by getting enterprises ready for Zero-Day Near Real-Time Scans.

Summary

We have cover 3 possible use cases for Memory Driven Computing, whereby core to the solution is a platform which supports Large Memory footprint, driving applications performance's several orders of magnitude to deliver unparalleled performance at attainable price points, for typical commercial grade businesses.

References

[1] 6-cybersecurity Mega Trends
https://www.hpe.com/us/en/insights/articles/6-security-megatrends-1905.html

[2] Scale-Up Graph Analytics with Reservoir Labs and Trovares:
https://www.reservoir.com/wp-content/uploads/2019/06/ENSIGN-SDFlex-WP.pdf

[3] Memory Drive Computing in Superdome Flex:
https://www.hpe.com/us/en/newsroom/blog-post/2017/05/memory-driven-computing-explained.html

[4] Mission critical Infrastructure for Data Driven Enterprise
https://www.hpe.com/hpe-external-resources/a00037000-7999/enw/a00037029?resourceTitle=Mission-critical+infrastructure+for+the+data-driven+enterprise