High Performance Data Analytics in Precision Medicine Using Scale-Up and Hybrid Supercomputing Solutions

October 2016

Authors

Michael S. Atkins
Director, Science Division

Life Science Teams

Genomic Pipeline
Tianwen Chu
Senior Data Scientist

Frank D'Ippolito
Senior Mathematician

Brian Furtaw
GPU Solutions Architect

Dave Wright
FCL Systems Administrator

Cancer Analytics
Terry Antony
Analytics Software Engineer

Sarah Bullard
Computer Scientist

Nicole Sabatelli
Bioengineer

Shawn Ulmer
Bioinformatician

Infectious Disease Analytics
Lucia Fernandez
Bioengineer

Kyle Milligan
Computational Biologist

Rui Ponte
Computer Programmer

Meena Sengottuvelu
Computer Scientist

Executive Summary

High Performance Data Analytics (HPDA) is an emerging technology that combines high performance supercomputing with advanced data analytics. The ultimate goal of HPDA in Precision Medicine is to provide medical professionals with crucial recommendations for the best course of treatment for individual patients based on all the relevant data available. FedCentric is developing scale-up and hybrid HPDA technologies to help medical professionals extract knowledge and insights from large, complex genomic databases to improve Precision Medicine for a number of diseases.

Scale-out supercomputing often involves linking together lower-performance machines to collectively do the work of a single scale-up machine. However, scale-out supercomputing has not proven to be cost-effective or provide the high-performance needed to analyze Precision Medicine data in real time. Scale-up supercomputing combines all the processing capacity and memory of a scale-out system in a single, more powerful machine. In scale-up supercomputing, massive datasets are analyzed completely in RAM at compute rather than network speed–minimizing latency issues–resulting in orders of magnitude speed and performance improvements. Hybrid supercomputing utilizes the unique advantages found in both scale-out and scale-up architectures, while also employing different processor types to accelerate speed and performance of different steps in a process. Hybrid supercomputing has the potential to be a disruptive technology in its ability to very quickly pipe or switch complex HPDA steps to the hardware best able to process it.

This white paper discusses hardware and HPDA solutions and some specific use cases FedCentric is developing to achieve Precision Medicine objectives in real time with exponentially increasing workloads. The Life Science Teams at FedCentric are currently working on scalable solutions in genomics and downstream analytics. Our Genomic Pipeline team has achieved significant speed improvements relative to recent benchmarks in the genomic sequence alignment and variant discovery pipeline using a hybrid supercomputing architecture. Our Downstream Analytics teams are developing graph and machine learning analytics to rapidly query post pipeline data like high quality genome and VCF files using the SGI UV300 scale up architecture.

Overview of Analytics Challenges in Precision Medicine

High Performance Data Analytics (HPDA) is an emerging technology that combines high performance supercomputing with advanced data analytics. HPDA is increasingly recognized by data scientists and researchers in academia, government, and industry as a leading technology to mine and extract actionable knowledge and insights from *Big Data* in real time to accurately inform decision makers. The ultimate goal of HPDA in Precision Medicine is to provide medical professionals with crucial recommendations for the best course of treatment for individual patients based on all the relevant data available.

High Performance Data Analytics (HPDA) is increasingly recognized by data scientists and researchers in academia, government, and industry as a leading technology to mine and extract actionable knowledge and insights from *Big Data* in real time to accurately inform decision makers.

A significant challenge to performing real time HPDA on contemporary compute (CPU) and data (I/O) intensive problems–like those in Precision Medicine–is that most organizations still use outdated legacy hardware or they simply don’t have the hardware needed to perform modern, complex data analytics. As a result, most organizations try pushing *Big Data* to more advanced cloud-based, scale-out servers at remote data centers to improve their ability to do HPDA while controlling and minimizing their own hardware, power, and maintenance costs. The biggest problem with this approach is latency: it is not a trivial matter to move *Big Data* to a cloud-based server over a network, particularly if the data size is increasing continuously. Many organizations are grappling with the latency of moving very large datasets to the cloud. When it comes to very *Big Data*, it is better to move analytics to the data rather than moving data to the analytics. And for certain types of data, scale-out clusters do not perform as well as scale-up and hybrid systems.

To better understand HPDA in Precision Medicine using scale-up and hybrid supercomputing solutions, we first need to define (and in some cases redefine) what terms like *Big Data*, Precision Medicine, and scale-out, scale-up, and hybrid supercomputing mean in the context of this white paper.

Big Data

Big Data has become an integral part of the global economy in every industry and every sector. Organizations create a tremendous amount of digital data simply as a by-product of normal business activities. Most large- and many mid-cap organizations generate terabytes, petabytes, and even exabytes of data each year about their customers, suppliers, operations, and from connected products. Billions of networked sensors have been embedded in mobile devices, appliances, and machines that sense, create, and communicate data as part of the Internet of Things. This is especially true in medical research, genomics, health care, and other biomedical organizations involved in Precision Medicine, with medical devices, DNA/RNA sequencing machines, and numerous other devices generating *Big Medical Data* continuously.

As a particularly relevant example, high throughput and next-generation sequencing have significantly increased the quantity of raw and processed genome sequencing data researchers need to process. High throughput sequencing technologies are producing data in such copious quantities that the cost of sequencing a base has decreased much faster than the cost of storing a byte. To compound matters, sequencing data is routinely stored in redundant sets as researchers process and annotate data iteratively and seldom, if ever, delete anything. As a result, researchers are struggling to handle massive volumes of genome sequencing data that are steadily increasing.

The term *Big Data* is best defined as datasets whose size is beyond the ability of traditional hardware and software tools to capture, store, manage, and analyze.

The term *Big Data* is best defined as datasets whose size is beyond the ability of traditional hardware and software tools to capture, store, manage, and analyze. This definition is intentionally subjective and dynamic: *Big Data* is not defined as being larger than a certain number of petabytes or exabytes. As technology changes through time, the size of datasets considered to be *Big Data* will also change.

Big Data may also vary by industry and sector, depending on the hardware and software tools used and the size of datasets common within a particular sector. *Big Data* in most sectors can range from terabytes to exabytes. While the use of *Big Data* will matter across sectors, some sectors are poised for greater gains. This white paper is about the biomedical, healthcare, and pharmaceutical sectors–the three largest sectors in Precision Medicine–that already have some of the biggest *Big Data* analytics challenges.

Precision Medicine

In his 2015 State of the Union address, President Obama launched the Precision Medicine Initiative to revolutionize how we improve health and treat disease. Until recently, most medical treatment models have been designed for the "average patient." This "one-size-fits-all" approach results in treatments that are successful for some patients but not for others.

Precision Medicine is an innovative medical treatment model that takes into account individual differences in patients' genomes, environments, and lifestyles. It provides doctors and medical researchers the ability to customize healthcare—with medical decisions, practices, and/or products– for individual patients.

Advances in Precision Medicine have already led to powerful new discoveries and several new treatments that are tailored to specific characteristics, such as a person's genetic makeup, or the genetic profile of an individual's tumor, that improve chances of survival and reduce exposure to adverse effects.

FedCentric is developing scale-up and hybrid HPDA technologies to help medical professionals extract knowledge and insights from large, complex genomic databases to improve Precision Medicine for a number of diseases.

FedCentric is developing scale-up and hybrid analytics technologies to help medical professionals extract knowledge and insights from large, complex genomic databases to improve Precision Medicine for a number of diseases.

For example, in-memory graph analytics is a very powerful tool for cancer researchers to determine relationships between genetic variants and specific types of cancer.

Traditional analytic approaches currently used by researchers are too slow to reveal useful information. FedCentric's proprietary HPDA using scale-up and hybrid supercomputing solutions enable medical professionals to visualize and mine connected data to develop and implement Precision Medicine in ways that are orders of magnitude faster and more intuitive than traditional scale-out approaches alone.

Scale-Out, Scale-Up, and Hybrid Supercomputing

Scale-Out Supercomputing

Scale-out (aka distributed or cluster) supercomputing is a traditional approach utilizing hundreds or even thousands of commodity servers connected in parallel through high-bandwidth networks. Well known examples of scale-out supercomputing include manufacturers like HP and Dell, cloud-based platforms like Amazon Web Services, Microsoft Azure, and Google Cloud, cluster distribution and processing software like Apache Hadoop, and network connections like Infiniband.

Scale-out supercomputing parallelizes *Big Data* problems through data partitioning to independent processing nodes in the cluster. Data parallelization allows analytics to be applied independently to each data partition in a dataset, which theoretically allows the degree of parallelization to *scale out* with the volume of data.

However, scale-out parallelization has not proven to be cost-effective or provide the high-performance needed to analyze very *Big Data* in real time, due to some fundamental limitations:

Scale-out supercomputing often involves linking together lower-performance machines to collectively do the work of a single scale-up machine.

- Scale-out computing often involves linking together lower-performance machines to collectively do the work of a single, more powerful scale-up server with significantly more processing capacity and RAM.
- Data and processing must be distributed in scale-out systems via network connections, which results in high latency with rapidly expanding workloads that need to be processed in real time.
- A good example of this in Precision Medicine is the high-throughput sequencing pipeline, which has latency bottlenecks in pushing massive genomic datasets to clouds and distributed clusters at network speed.
- Latency will become an intractable issue when Precision Medicine objectives require millions of individual genomes to be sequenced and analyzed in meaningful time frames to yield actionable information.

IDC[1] summed up the benefits and limitations of scale-out computing this way:

> *“Scale-out refers to expanding to multiple servers rather than a single bigger server. The use of availability and clustering software (ACS) and its server node management, which enables IT managers to move workloads from one server to another or to combine them into a single computing resource, represents a prime example of scale out. Scale out usually offers some initial hardware cost advantages (e.g., four 2-socket servers may cost less than a single 16-socket server). (IDC notes that each socket may have more than one processor.) In many cases, the redundancy offered by a scale-out solution is also useful from an availability perspective. However, IDC research has also shown that scale-out solutions can drive up opex to undesirable levels. At the same time, large data volumes and required processing capabilities are taxing scale-out systems.”*

Scale-out parallelization has not proven to be cost-effective or provide the high-performance needed to analyze very *Big Data* in real time, due in large part to the latency of processing massive datasets at network speed.

Scale-Up Supercomputing

Scale-up supercomputing combines all the processing capacity and memory of a scale-out system in a single, more powerful machine. This is a useful way to scale databases and a number of other *Big Data* workloads without having to distribute them at slower network speed to smaller nodes in a cluster. In scale-up supercomputing, massive datasets are analyzed completely in RAM at compute rather than network speed–minimizing latency issues–resulting in orders of magnitude speed and performance improvements. Examples of scale-up systems include SGI UV3, IBM Power Enterprise servers, and Oracle SPARC M7-16.

In scale-up supercomputing, massive datasets are analyzed completely in RAM at compute rather than network speed–minimizing latency issues–resulting in orders of magnitude speed and performance improvements.

One particularly useful feature of scale-up machines is that they can be configured to perform like a scale-out system with all the advantages of distributed supercomputing, but without the latency issues. The converse is not true: scale-out cannot be configured to perform as a scale-up system.

Some limitations to consider with scale-up supercomputing are:

- Scale-up machines can be more expensive than a comparable scale-out configuration.
- There is a physical limitation to the computing power and memory you can have in a single machine.
- Most software is optimized for scale-out supercomputing.
- Some analytics problems are highly parallelizable and simply run better in a scale-out system.

IDC[1] summarized the value proposition of scale-up supercomputing as follows:

> *"Scale-up server solutions not only represent a comparable alternative to scale-out or distributed environments but also, for a variety of reasons, can have a demonstrably beneficial impact on an organization's ability to gain insights and compete more successfully. More and more organizations — from small and medium-sized businesses to the largest enterprises — are deploying compute- intensive and data-intensive applications to analyze huge data sets so as to obtain new knowledge and insights and build innovative products or marketing strategies around them. Quite a few of these applications have traditionally resided on HPC clusters, and others have been deployed in distributed environments. But the limitations of clusters to perform analytics on very large data sets in-memory and scale without severe latency implications caused by the cluster networking have led to a revived interest in scale-up platforms.*
>
> *An IDC study on the business value of scale up shows that scale up can yield improved performance, including increased resource utilization, improved application performance, less unplanned downtime, and extended datacenter life. Furthermore, we found that scale-up consolidation can result in combined savings of 33% thanks to operational cost reductions as a result of reduced server management staffing requirements and lower costs associated with power and cooling, software licensing, and IT infrastructure."*

"The limitations of clusters to perform analytics on very large data sets in-memory and scale without severe latency implications caused by the cluster networking have led to a revived interest in scale-up platforms."

Hybrid Supercomputing

Hybrid supercomputing utilizes the unique advantages found in both scale-out and scale-up architectures, while also employing different processor types (e.g., CPU, GPU, FPGA, MIC, ARM) to accelerate speed and performance of different steps in a process. FedCentric is pioneering the development of hybrid supercomputing solutions for *Big Data* problems with multiple steps and processes that do not all run optimally in a single type of compute environment. For example, in the genomic pipeline some steps (e.g. mapping to reference, dedupping) work better with scale-out parallelization using CPUs and FPGAs while other steps (e.g. de novo assembly, joint genotyping, downstream analysis) benefit from a single, large memory scale-up system using GPUs. A hybrid supercomputing configuration offers the best of scale-out and scale-up.

Hybrid supercomputing utilizes the unique advantages found in both scale-out and scale-up architectures, while also employing different processor types (e.g., CPU, GPU, FPGA) to accelerate speed and performance of different steps in a process.

Hybrid supercomputing has the potential to be a disruptive technology in its ability to very quickly pipe or switch complex HPDA steps to the hardware best able to process it. FedCentric uses hybrid commodity hardware (SGI UV3 & ICE), open source software (Apache Spark), and customizable HPDA in a modular appliance at an affordable price point for small organizations that can scale to handle the largest *Big Data* and HPDA problems for very large organizations.

Putting It All Together: HPDA in Precision Medicine Using Scale-Up and Hybrid Supercomputing Solutions

HPDA in Precision Medicine has multiple distinct phases, each of which introduces unique challenges. These phases can include data acquisition, cleaning, distribution, modeling, parsing (structured and unstructured data types), ingestion, compute and data intensive analysis, querying, and interpretation. Many data scientists and researchers focus only on a single phase of interest without considering other crucial aspects of the entire process or they may not know how to approach phases outside their area(s) of expertise. Some useful questions to ask before starting an HPDA project are:

HPDA in Precision Medicine has multiple distinct phases, each of which introduces unique challenges.

- ➢ *How to assess the quality of initial input data and the accuracy of final output data?* This often requires interdisciplinary teams of subject matter experts (SMEs) in biomedical research fields working with mathematicians, data scientists, and hardware engineers to ensure that input data is of sufficient quality to yield meaningful results and that the entire process is able to produce reliably accurate output data, which must be validated by SMEs. As we all know, garbage in equals garbage out. There must be validation steps along the way done by competent experts.

- ➢ *What kinds of queries do we want to perform?* Queries can be simple or complex and they typically will not be apparent in advance. SMEs may need to determine the right queries to ask based on the data they are using. Doing this requires smarter solutions and better support for user interaction with the HPDA pipeline. In Precision Medicine, there is a major bottleneck in the number of people able to query data properly and analyze it. This number can be improved by supporting many levels of engagement with the data, not all requiring deep database and/or HPDA expertise.

- ➢ *What is the best way to parse structured and unstructured data types?* SMEs determine the different data types they want to query and data often comes in different formats and is not always in a format ready for analysis. Biomedical data of interest to SMEs can be very diverse such as DNA/RNA sequence information, variant/mutation calls, clinical data, health records, sensor data from patient monitoring systems, image data from x-rays, MRIs, and CT scans, and other data in a multitude of format types, both structured and unstructured. Data scientists work directly with SMEs to parse together diverse data types to better ensure useful output.

Subject matter experts (SMEs) determine the different data types they want to query and data often comes in different formats and is not always in a format ready for analysis.

- ➢ *Which database technology is best for different data types and queries?* There are several different database technologies out there, each with unique advantages and disadvantages for HPDA in Precision Medicine. The best database technology will depend on several interrelated factors including data types and queries, but also on the types of HPDA to be performed on the database, such as machine learning, phylogenetic analysis, etc. Database technologies are discussed in more detail below.

- *What hardware is optimal for the analytics steps in a pipeline?* Even though most researchers are pushing their data to cloud-based services, this does not mean that scale-out cloud servers are always the best hardware choice for doing complex HPDA. If speed and performance are mission critical, the latency of moving Big Data at network speed to a cloud server or between nodes in a cluster will be a problem. If patient data privacy is a concern, sending and storing sensitive data offsite may not be the right solution. Each HPDA problem is unique and the optimal hardware will be determined by the project objectives and requirements.

- *Will HPDA in Precision Medicine always result in accurate recommendations for the best course of treatment for individual patients?* Nothing is perfect, and HPDA is no exception. There are many reasons why we should not expect the output of HPDA in Precision Medicine to be 100% accurate, most of which come down to technical and human errors in data collection as well as other issues. For example, patients may choose to hide risky behavior and caregivers may sometimes misdiagnose a condition. Patients may inaccurately recall the name of a drug or even that they ever took it, leading to missing information in the history portion of their medical record. Existing work on data cleaning assumes well-recognized constraints on valid data or well understood error models. For HPDA in Precision Medicine these may not exist yet and will certainly influence some portion of the results.

Precision Medicine is a decades-long endeavor to improve individual patient treatments using *Big Data* and HPDA.

Precision Medicine is a decades-long endeavor to improve individual patient treatments using *Big Data* and HPDA. It will take time to fine tune all the phases of this process to achieve the objectives set forth in 2015.

FedCentric is addressing these kinds of questions and the overall objectives of the Precision Medicine Initiative by creating a complete and comprehensive solution that combines the latest in scale-up and hybrid supercomputing with advanced HPDA to get meaningful knowledge and insights from *Big Data* in Precision Medicine. Following are detailed examples of the kinds of HPDA technology and some specific use cases we are developing to achieve Precision Medicine objectives in real time with exponentially increasing workloads.

Apache Spark: A Big Data Processing Framework

Apache Spark is an open source big data processing framework built around speed, ease of use, and sophisticated analytics. It was originally developed in 2009 in UC Berkeley's AMPLab, and open sourced in 2010 as an Apache project. Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data scientists to efficiently execute streaming, machine learning and other workloads that require fast iterative access to datasets. It works in scale-out, scale-up, and hybrid supercomputing environments.

Performing compute and data intensive analytics on large, diverse biomedical data requires a data processing framework that optimizes efficient use of the underlying hardware capability. Apache Spark is that framework.

Performing compute (CPU) and data (I/O) intensive analytics on large, diverse biomedical data requires a data processing framework that optimizes efficient use of the underlying hardware capability. Apache Spark is that framework. For example, FedCentric's SGI UV300 system has 256 cores that work with Spark to use all the available core. Spark essentially allows a scale-up system, with all of its advantages, to also exploit many of the benefits of a scale-out system by introducing a new data structure called the Resilient Distributed Dataset. This data structure allows large amounts of data to be operated on in parallel by Spark's virtual executor nodes, under the management of a single fat virtual driver node. Since the nodes are virtual, internodal communication is instant. Each core acts as an executor with an independent java heap, amounting to 256 executors operating in parallel. This architecture gives us the ability to scale processes up to 256x faster than a single threaded system.

Database Technologies

One of the biggest challenges in analyzing massive data sets is identifying significant relationships at a granular level. Finding important signals in latent, diverse data sets helps data scientists better understand and explore unique relationships connecting individual data points. However, massive data sets are typically stored in asymmetric databases that are intractable to traditional approaches, such as relational database and statistical analytics.

Traditional relational databases are often not optimized for the complex queries Precision Medicine researchers want to perform. The data is typically found in complex structured and unstructured databases that are not well suited for relational databases. While structured query language (SQL) and relational databases are well-established, easily understood, and standardized, they are not ideal to handle the volume or diversity of Precision Medicine data that is available. In order to incorporate novel, variable data into such a database, different tables must be created and joined across various keys, which is useful in storing data, but fails to efficiently query data across tables and keys. Furthermore, relational databases are not well-suited to the development of an efficient and intuitive tool for researchers, due to the size, complexity, and diversity of the data that needs to be incorporated.

Graph analytics is the science of creating graph databases and algorithms that target granularity and weigh how individual data objects are related to each other.

Determining how data objects relate to each other and their patterns of connections is more important than simply classifying and summarizing them.

Graph analytics is the science of creating graph databases and algorithms that target granularity and weigh how individual data objects are related to each other. Determining how data objects relate to each other and their patterns of connections is more important than simply classifying and summarizing them. Graph analytics allow users to visualize and mine connected data in ways that are orders of magnitude faster and more intuitive than traditional methods.

Graph databases are a robust alternative to relational tables and use a node and edge structure for semantic queries. Nodes represent subjects like *Person* or *Company*, and edges represent relationships like *employedBy*. To observe the relationship between two subjects in a relational structure, you need to join two subject tables to a relationship table. In contrast, a graph structure performs the same action through edge traversal, which is more efficient.

There are two types of graph databases, RDF triples and property graphs. Resource description framework (RDF) triples maintain some of the structure of the relational database, but are overall more flexible. Each triple consists of a subject node, an object node, and a predicate edge that joins the two. RDF triples are rigid and tend to result in an increase in the size of the database. Property graphs also have nodes and edges but allow information to be stored on the node and edge itself, reducing the size of the graph. The property graph is more flexible than RDF triples because it allows the creators to choose the best possible configuration of data. Instead of being forced to store information in triplets, property graphs are able to sustain an unlimited number of node-edge relationships. Join operations are not necessary to analyze dense, interconnected data in a property graph database. Querying the database merely accesses the list of relationship records, eliminating the need for a lengthy search and match computation. Graph databases do not require a pre-set schema and any amount of new data can be arbitrarily incorporated with ease. This flexible system allows researchers to efficiently recognize patterns within such data.

Graph models are gaining popularity as a better technology to analyze complex, heterogeneous data and could prove essential in the development and implementation of Precision Medicine.

Graph models are gaining popularity as a better technology to analyze complex, heterogeneous data and could prove essential in the development and implementation of Precision Medicine. The sheer size and density of graph databases had led researchers to dismiss the practicality of graph models until very recently. As a result, researchers have relied for years on traditional relational databases such as MySQL and PostgreSQL. Unfortunately, SQL databases are not suited for graph models and are proving to be a limitation on researchers' ability to find unique relationships in diverse biomedical data.

Machine Learning, Deep Learning, and Neural Networks

Machine learning is a method of data analysis that automates analytical model building and is used in predictive analytics or predictive modeling. Machine learning explores the construction and study of algorithms that can learn from and make predictions from massive data sets. Machine learning algorithms operate by developing models from example inputs to make data-driven predictions or decisions, rather than following strictly static program instructions.

Machine learning algorithms operate by developing models from example inputs to make data-driven predictions or decisions, rather than following strictly static program instructions.

Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results.

Machine learning algorithms fall into three categories: supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning is when all the input data has an output label and the prediction algorithms learns through the accuracy of its prediction. The two main types are classification and regression. A classification problem is when the output is a category and regression problem is when the output is a value. For example, the Infectious Disease Analytics Team at FedCentric uses a support vector machine, a classification algorithm, to classify viruses based on data derived from their gene sequence.

Unsupervised learning is when all the input data is unlabeled and the algorithm creates its own predictions based on relationships within the data. The two types are clustering and association. Clustering algorithms group data points together based on certain attributes. Association algorithms attempt to define rules and implications based on the data. The Infectious Disease Analytics Team uses hierarchical clustering to group viral sequences based on distance calculations for quick phylogenetic tree creation.

Semi-supervised Learning is when only a portion of the input data is labeled and the algorithm uses a combination of supervised and unsupervised techniques for prediction. Many real world machine learning problems fall into this category because data is often incomplete. Unsupervised learning can be used to predict the structure of the data, while supervised learning can be used to predict missing data. Overall, machine learning is used to explore the links between data to derive insight.

Deep learning (aka deep structured learning, hierarchical learning, or deep machine learning) is a branch of machine learning based on a set of algorithms that attempt to model high level abstractions in data by using a deep graph with multiple processing layers, composed of multiple linear and non-linear transformations.

Deep learning is part of a broader family of machine learning methods based on learning representations of data.

Deep learning is part of a broader family of machine learning methods based on learning representations of data. An observation (e.g., an image) can be represented in many ways such as a vector of intensity values per pixel, or in a more abstract way as a set of edges, regions of particular shape, etc. Some representations are better than others at simplifying the learning task (e.g., face recognition or facial expression recognition). One of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction.

A neural network is a system of hardware and/or software patterned after the operation of neurons in the human brain. Neural networks -- also called artificial neural networks -- are a variety of deep learning technologies. Commercial applications of these technologies generally focus on solving complex signal processing or pattern recognition problems. A neural network usually involves a large number of processors operating in parallel and arranged in tiers. The first tier receives the raw input information -- analogous to optic nerves in human visual processing. Each successive tier receives the output from the tier preceding it, rather than from the raw input -- in the same way neurons further from the optic nerve receive signals from those closer to it. The last tier produces the output of the system.

Each processing node has its own small sphere of knowledge, including what it has seen and any rules it was originally programmed with or developed for itself. The tiers are highly interconnected, which means each node in tier n will be connected to many nodes in tier n-1 -- its inputs -- and in tier n+1, which provides input for those nodes. There may be one or multiple nodes in the output layer, from which the answer it produces can be read.

Neural networks are notable for being adaptive, which means they modify themselves as they learn from initial training and subsequent runs provide more information about the world.

Neural networks are notable for being adaptive, which means they modify themselves as they learn from initial training and subsequent runs provide more information about the world. The most basic learning model is centered on weighting the input streams, which is how each node weights the importance of input from each of its predecessors. Inputs that contribute to getting right answers are weighted higher.

Precision Medicine Use Cases

Genomic Pipeline

FedCentric's Genomic Pipeline Team has achieved significant speed improvements (10x-100x) relative to recent benchmarks[2,3] in the genomic sequence alignment and variant discovery pipeline using the SGI UV300 scale-up architecture. Since most algorithms and code were written for scale out configurations, it has been a challenge for us to optimize them for scale up, but there is additional progress to be made by optimizing specific steps that we think will work better in a single, large memory instance such as de novo assembly and joint variant discovery, as well as with large sequence files over 100GB. Based on benchmark results from scale-out and scale-up systems, FedCentric is developing a hybrid supercomputing system that will switch steps in the pipeline to hardware best able to process output data from a previous step.

Aligning hundreds of thousands to millions of individuals' genome sequencing output to a reference genome is a highly parallelized process that is optimized for scale-out supercomputing using CPU and FPGA processors. Subsequent cleaning and other pre-processing steps also run better on scale-out architectures. However, de novo assembly of sequencing reads utilizes De Bruijn graph technologies to improve assembly results, and sharding graphs across a cluster of scale-out machines significantly decreases speed and performance. De novo assembly may be much faster on a scale-up machine using GP-GPU processors. Joint genotyping in the variant discovery part of the pipeline is improved by comparing multiple genomes together against the reference genome. The larger the file size that must be analyzed together, the better the fit for a single, large memory scale-up machine.

The hybrid supercomputing approach will greatly improve speed and performance throughout the genomic pipeline and in downstream analysis of pipeline output data.

The hybrid supercomputing approach will greatly improve speed and performance throughout the genomic pipeline and in downstream analysis of pipeline output data.

Cancer Analytics

The Cancer Analytics team is developing a large graph database of heterogeneous cancer data with advanced machine and deep learning algorithms to quickly find unique, potentially important relationships in complex data. The past several years have seen a significant increase in high-throughput experimental studies that catalog variant datasets using massively parallel sequencing experiments. New insights of biological significance can be gained by this information with multiple genomic location based annotations. However, efforts to obtain this information by integrating and mining variant data have had limited success so far and there has yet to be an efficient method developed that can be scalable, practical and applied to millions of variants and their related annotations. Relational databases have demonstrated to be capable of handling these tasks, but have proven to be an inflexible tool for analysis of variant data.

The representation of data within graph data structures demonstrates promise in determining which specific genomic locations are highly correlated with phenotypic outcomes or even specific variables.

In Phases I & II of a 2015-16 collaboration with the Frederick National Laboratory for Cancer Research (FNLCR), FedCentric implemented a graph database on an SGI UV300 scale-up system to compare query speed and performance against a relational database on a scale-out system at FNLCR. Specifically, we explored the application of graph data structures as a proof of concept for scalable interpretation of the impact of variant related data. Our results suggest that graph databases can reduce the time and effort needed to perform more advanced computational analytics upon variant data. The representation of data within graph data structures demonstrates promise in determining which specific genomic locations are highly correlated with phenotypic outcomes or even specific variables. We also investigated how graph databases can be used to analyze high-density genomic datasets.

In the first two phases, we used variant and population data from the 1000 genomes project and annotation data sets from EntrezGene, ClinVar, and UniProt. A graph model was developed using Sparksee technology to best answer the queries on variant data and assess performance. All the data sets were parsed and the graph was created with nearly 200 million nodes and 12 billion edges. In Phase I of the project, the graph was used for retrieval of all relevant information for single variants. In Phase II, we explored complex queries for identifying clusters and patterns within individuals based on their differences with respect to the reference genome.

Our results showed that for single, variant-based annotation searches, the speed and performance of the graph model on a scale-up system was similar to a well-architected relational database. However, the graph model allowed us to test complex queries for identifying variant patterns that we were unable to run using the relational architecture in scale-out. In particular, graph in a scale-up system easily handled clustering methods, which were too computationally intensive for relational databases on FNLCR's scale-out machines. By classifying different sorts of variants into groups based on specific genetic variables, cancer researchers will be able to see complex relationships among and between genetic data.

Highlights of FedCentric's 2015-16 collaboration with FNLCR:

- Loaded genomic data from 2,500 individuals and 10 major genomic databases into a graph model.
- Developed a domain specific language for the graph model that is user friendly.
- Developed advanced queries suggested by FNLCR researchers to fully explore the potential of graph database technology.
- Demonstrated the ability of graph technology to perform complex clustering algorithms and deliver unique insight to researchers.

The result is not only descriptive analytics that yield information about populations, individuals, genes, and variants, but through a layered combination of graph and machine learning for example, we can deliver predictive analytics to better inform researchers in Precision Medicine decisions.

In Phase III of this work, the team is expanding the graph database to contain cancer data from The Cancer Genome Atlas (TCGA). We incorporated a variety of new data types including variant data, gene expression, miRNA, DNA methylation, clinical data, and copy number variation from over 11,000 cancer patients. Data was parsed and standardized from a diversity of file types including XML, TXT, VCF, TSV, and CSV. This new data added over 119 million nodes and 3.2 billion edges to the Phase II graph.

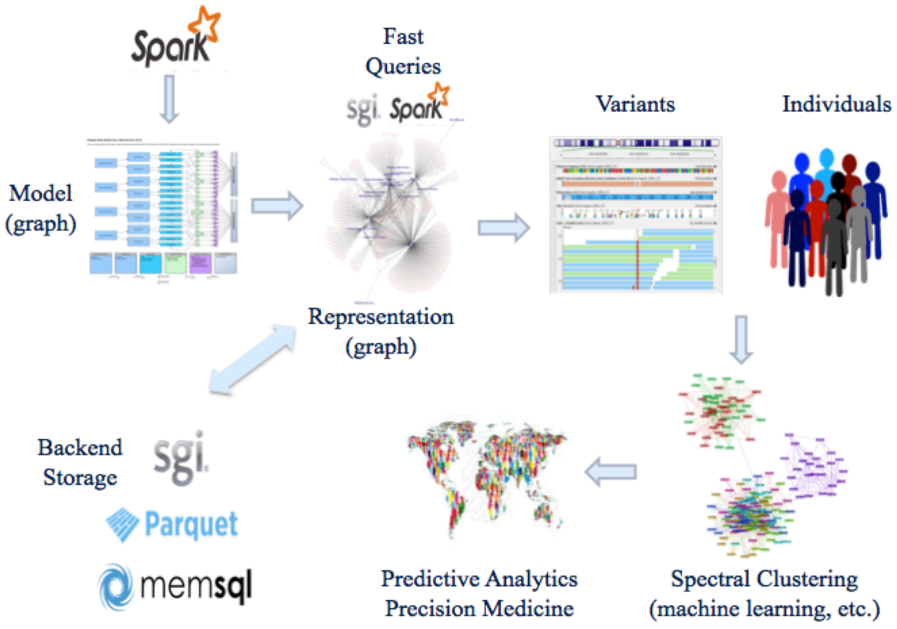

Figure 1. Workflow in Cancer Analytics. Cancer researchers can quickly query variants and individuals that apply spectral clustering (machine learning) algorithms to groups of people based on their genetic variation. The result is not only descriptive analytics that yield information about populations, individuals, genes, and variants, but through a layered combination of graph and machine learning for example, we can deliver predictive analytics to better inform researchers in Precision Medicine decisions.

Infectious Disease Analytics

The Infectious Disease Analytics Team is applying HPDA and scale-up supercomputing to data from infectious diseases to develop a robust epidemiological tool that can be used to quickly identify, monitor, and contain outbreaks wherever they occur, from remote hot zones in Africa to a nosocomial infection in a local hospital. We are developing it using data on Flaviviruses sequences. The flavivirus genus consists of 73 mosquito- and tick-borne viruses that pose a considerable threat to public health, including Dengue virus, West Nile virus, and Zika virus.

All flaviviruses are encoded by a single-stranded, positive sense RNA genome roughly 11 kb in length. The genome consists of 10 proteins: the capsid (C), membrane (M), and envelope (E) perform structural functions and NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5 perform nonstructural functions. A deeper understanding of the molecular evolution of flaviviruses is needed to guide public health decisions and to prevent flavivirus epidemics, such as the 2015-2016 Zika virus epidemic in the Americas. Comparative genomic analysis can lead to many discoveries regarding molecular evolution and epidemiology, but computational power remains a limiting factor due the size of genomic data. Genomics is a big data science and is considered a “four-headed beast” because it is demanding in terms of data acquisition, storage, distribution, and analysis.

FedCentric is applying HPDA and scale-up supercomputing to data from infectious diseases to develop a robust epidemiological tool that can be used to quickly identify, monitor, and contain outbreaks wherever they occur, from remote hot zones in Africa to a nosocomial infection in a local hospital.

The Infectious Disease Analytics Team is developing the Infectious Disease Epidemiology Appliance (IDEA) for the rapid analysis and visualization of over 6,200 genomic and 93,000 proteomic sequences. IDEA allows users to select from a list of flavivirus genomic and proteomic sequences or upload their own FASTA files to analyze. The database’s capabilities include calculation of local sequence alignments, that is regions of similarity between two nucleotide or protein sequences, generation of phylogenetic trees and epidemiological maps, and rapid querying of viral data.

IDEA is able to rapidly generate possible phylogenetic trees of flaviviruses and their individual proteins. IDEA allows for the comparison of genomic and proteomic flavivirus sequences by employing the Maximum Likelihood algorithm. Maximum Likelihood methods involve using statistical techniques to infer a probability distribution to describe the given sequence data. The Likelihood of a tree is proportional to the probability of observing the data given the associated probability distribution. The tree selected is the one that maximizes this parameter given the alignment and sequence data. The algorithm evaluates a branching pattern in terms of the probability that it would have lead to the associated sequence data. Maximum Likelihood methods are generally considered to be more accurate than distance based methods, and are considered more thorough than Neighbor Joining methods as well because they allow for the evaluation of multiple tree topologies.

It is also necessary to select a base substitution model in order to calculate the likelihood for a given data set. The Jukes Cantor model allows for forward and backward mutation at specific DNA sites, and uses equal probabilities of substitution between all base types. This simple model was determined to be the best for IDEA, but another one that assigns different probabilities to transitions and transversions can be easily implemented. Confidence in the phylogenetic tree is then assessed via the statistical bootstrap method. Using this method a matrix of taxa x characters is sampled with replacement to create many new matrices of the same size as the original. These are then used to find the best-fit tree by assigning confidence values to internal branches of the tree. One of the biggest challenges with Maximum Likelihood methods is that they are very computationally expensive. The algorithm is classified as NP-Hard, putting it among the toughest computer science problems to solve efficiently. Scale-up supercomputing allows researchers to significantly speed up this calculation, quickly producing accurate phylogenetic trees for real time analysis and tracking of flavivirus strains.

Scale-up supercomputing allows researchers to significantly speed up this calculation, quickly producing accurate phylogenetic trees for real time analysis and tracking of flavivirus strains.

An epidemiological map depicting flavivirus cases from the genomic data allows IDEA to track the spread of flaviviruses over space and time. Each viral sequence in the database is accompanied by the date and location where it was retrieved. IDEA plots these locations to a world map, and a slider allows the user to visualize the spread of the virus over time. This feature is only available for viruses with an adequate number of samples.

IDEA offers the ability to query the database for information specific to the user's needs. In addition to the likelihood calculations used in the tree generator, Levenshtein distances between sequences can be calculated and displayed, and the data can be sorted by any feature included in the VIPR database (date, location, sequence type, etc.). IDEA can also export data in PNG, CSV, or FASTA format.

IDEA will be an invaluable tool to virologists, epidemiologists, and public health officials. The ability to accurately decipher the origin of a pathogenic strain will allow for preventative interventions and quarantines when they are needed most, not after the fact. This technology can be rapidly expanded to other pathogenic organism as well, such as the Ebola virus, HIV, Malaria, influenza, and even bacterial pathogens.

References

(1) IDC (2016). *The value proposition of scale-up x86 servers: cost efficient and powerful performance for critical business insights*. Document #US41083516. www.idc.com

(2) Diao, Y., Roy, A., and Bloom, T. (2015). *Building highly optimized, low-latency pipelines for genomic data analysis*. 7th Biennial Conference on Innovative Data Systems Research (CIDR '15), January 4-7, 2015, Asilomar, California, USA.

(3) Dandapanthula, N. and Yoon, K. (2015). *Dell HPC System for Genomics v2.0. A Dell Reference Architecture. HPC Infrastructure for Next Generation Sequencing Analysis*. Dell HPC Engineering. http://en.community.dell.com/techcenter/blueprints/blueprint_for_hpc/m/mediagallery/20441607/download

Important Note

In November 2016, the purchase of SGI by Hewlett Packard Enterprise (HPE) will be completed. This purchase underscores and validates the increasing importance of scale-up and hybrid supercomputing for complex HPDA. FedCentric has been a long-term Platinum Partner with SGI and is a new Top-Tier Partner with HPE in scale-up and hybrid supercomputing.

About FedCentric

FedCentric Technologies develops High-Performance Data Analytics (HPDA), graph database analytics, and machine learning for several applications and industries including network security, fraud prevention, and precision medicine. FedCentric's HPDA tools provide our clients with new opportunities to discover data-driven impacts and competitive advantages over traditional approaches.

FedCentric Technologies, LLC
4511 Knox Road, Suite 300
College Park, MD 20740
301.263.0030
www.fedcentric.com