BIG DATA ANALYTICS AND ITS LIKELY APPLICATIONS IN DEFENCE AND SECURITY FORCES ORGANISATIONS

AUTHOR BY | AIR CMDE T CHAND (RETD)

"You are entitled to your own opinion, but you are not entitled to your own facts**"** Daniel Patrick Moynihan

Big data analytics (BDA) deals with the meaningful analysis of very large volumes of data, also called big data. The big data is accumulated from a number of sources, such as social media networks, scientific data, intelligence work and inputs from various other sensors. The aim of analysis of this large data is to reveal patterns and trends that might otherwise be invisible. Through this knowledge competing organisations may be able to gain an advantage over their competitors and make superior decisions. Software tools such as Hadoop, MapReduce and NoSQL are used for the analysis. BDA applications are mainly in scientific, business and intelligence world. Applications for defence and security forces have also emerged and are evolving rapidly. In this paper, BDA has been explained; its framework and applications in general and likely applications for defence and security forces organisations in particular have been suggested.

Big Data

A widely accepted definition of the Big Data was formulated by Gartner who updated it in 2012 stating that "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimisation." This definition was expanded later to include other complementary characteristics such as machine learning. This type of data requires a different processing approach called big data analytics, which uses massive parallelism on readily-available hardware.

Big Data Analytics

Data is required to be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Analysis of data, also known as data analytics, is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information,

suggesting conclusions, and supporting decision-making. Software for data analysis is freely available . DevInfo is a database system endorsed by the United Nations Development Group (UNDG) for monitoring and analysing human development. ELKI is a Data mining framework in Java with data mining oriented visualisation functions. KNIME is the Konstanz Information Miner, a user friendly and comprehensive data analytics framework. PAW is a FORTRAN/C data analysis framework developed at CERN (European Organisation for Nuclear Research). Orange is a visual programming tool featuring interactive data visualization and methods for statistical data analysis, data mining, and machine learning. R is a programming language and software environment for statistical computing and graphics. ROOT is a C++ data analysis framework developed at CERN.

Sophisticated software programmes are used for big data analytics, but the unstructured data used in big data analytics may not be well suited to conventional data warehouses. Big data's high processing requirements may not be met by traditional data warehousing. As a result, newer, bigger data analytics environments and many technologies have emerged. These technologies have developed open-source software framework for processing large data sets over clustered systems.

Open-source big data analytics generally refers to the use of open-source software and tools for analysing large quantities of data for gathering relevant information by an organisation. The main open-source software employed for big data analytics is Apache's Hadoop. It is the most widely used software library for processing enormous data sets across a cluster of computers using a distributed process for parallelism. Many other components are required before a real analytics system can be formulated. Hadoop is the open-source implementation of the MapReduce algorithm pioneered by Google and Yahoo and is the basis of most analytics systems today. Open-source big data analytics services encompass: data collection system, control center for administering and monitoring clusters, machine learning and data mining library, application coordination service, and compute engine and execution framework.

Contribution by Technology Companies

Dr. Kulshrestha, suggests that technological developments have led to a synergetic relationship with digital industry where in the military no longer develops its own hardware and software, but harnesses and modifies the ‘commercial of the shelf’ (COTS) items. Common technologies in the big data ecosystem such as Apache Hadoop, Apache Hive / Apache Pig, Apache Sqoop, In-memory Databases, NoSQL Databases and MPP Platforms are being used by the defence forces globally and technology companies have developed systems for display of common operational picture for the defence forces. Several softwares have been developed for this purpose. Modus Operandi, accepts big data, and creates a common framework for easy identification of patterns.

Palantir Gotham developed by Palantir Technologies is used by the US Intelligence Community and counterterrorism analysts. Oracle has created a newly engineered system to handle big data operations which could be used for defence applications. Leidos has developed a Scale2Insight (S2i) software supporting large complex data environments with multiple disparate sensors for collecting information on different parts of the data ecosystem. SYNTASA provides analytical applications which focus on the behavior of visitors to internal government websites to determine the behavioral trends of visitors in order to improve the use of information by government analysts.

Big Data Analytics Policy Framework and Applications

Big Data Initiative (BDI) is one of the scientific programmes of the Department of Science and Technology (DST), Govt of India. DST believes that one of the most significant application scenarios where big data is generated is scientific computing. Scientists and researchers produce huge amounts of data everyday through experiments. But extracting useful knowledge for decision making purposes from these massive, large-scale data repositories is almost impossible for actual DBMS (Data Based management System) inspired analysis tools. A new methodology is required for transforming big data stored in heterogeneous and different type's data sources such as; legacy systems, web, scientific data repositories, sensor and stream databases and also social networks into a structured and interpretable format for target data analytics. As a result, data-driven approaches, in biology, medicine, public policy, social sciences, and humanities, can replace the traditional hypothesis-driven research in science . Some of the science and technology challenges that researchers across the globe and as well as in India facing are related to data deluge pertaining to astrophysics, materials science, earth and atmospheric observations, energy, fundamental science, computational biology, bioinformatics and medicine, engineering and technology, GIS and remote sensing, cognitive science and statistical data. These challenges require development of advanced algorithms, visualisation techniques, data streaming methodologies and analytics. Main constraints are: storage and computational power of IT eco system; algorithm design, visualization, scalability (machine learning, network and graph analysis, streaming of data and text mining), distributed data, architectures, data dimension reduction and implementation using computer science; statistics, optimisation, uncertainty quantification, model development (statistical, basic simulation) analysis and systems theory using mathematical sciences and contextual problem solving using multi disciplinary approach. DST has identified important areas for development of BDA eco system in India. Creation of the talent pool is the first requirement. This will require creation of industry academia partnership to groom the talent pool in universities as well as development of strong internal training curriculum to advance analytical depth. Collaboration, capability development and value creation are a few

other steps. DST has outlined its broader contours for BDI programme for the next ten years in five steps. Firstly; to promote and foster big data science, technology and applications in the country and to develop core generic technologies, tools and algorithms for wider applications in Govt. Secondly; to understand the present status of the industry in terms of market size, different players providing services across sectors, SWOT of industry, policy framework and present skill levels available. Thirdly; to carry out market landscape survey for assessing the future opportunities and demand for skill levels in next ten years. Fourthly; to bridge the skill level and policy framework gaps. Lastly; to evolve a strategic road map and micro level action plan clearly defining roles of various stakeholders such as government, industry, academia and others with clear timelines and outcome for the next ten years.

National Data Sharing and Accessibility Policy (NDSAP) 2012 of DST is designed to promote data sharing and enable access to government owned data for national planning and development . Defence forces can also make use of this policy framework.

Big Data Analytics Infrastructure Framework and Applications

Big Data Analytics infrastructure development in India is being steered by the C-DAC (Centre for Development of Advanced Computing), Ministry of Electronics and Information Technology (MeitY). The centre recently announced the launch of a new system called eVidur for tracking social media data . The system will use BDA and natural language processing based solutions for sentimental analysis of social media data. This would make it possible to auto analyse and interpret social media big data to capture the sentiment of people. The eVidur system will be used to track comments and opinions in real time and could be used to track people's views on political or development agendas or policy initiatives by government to take corrective measures. State of the art hardware system and networking environment has been created by the C-DAC at its various facilities.

C-DAC HPC (High Performance Computing) supercomputing systems, PARAM Yuva II, the latest in the series, is a eight-core, dual-socket node based hybrid compute cluster with multiple interconnects, compute co-processor, hardware accelerator, high performance storage and supporting software for parallel computing. It incorporates C-DAC's in-house technologies including PARAMNet-3 - a high speed system area network. PARAM Yuva II has consistently given a performance of over 386 teraflops. It has high bandwidth storage of 200 terabytes. C-DAC has commissioned and operates three national supercomputing facilities for HPC users, namely; National PARAM Supercomputing Facility at C-DAC, Pune; C-DAC's terascale supercomputing facility at C-DAC, Bangalore and Bioinformatics Resources and Applications Facility (BRAF) at C-DAC, Pune. These facilities are used by scientists and researchers across the nation for carrying out their research effectively. While the first two facilities cater to the

requirements of users from different applications domains, the third facility is specifically for users working in computational biology domain .

The supercomputing systems and facilities of C-DAC are used to solve computationally challenging problems in a number of areas of practical significance. These include: computational atmospheric science (mesoscale modeling, climate system model, medium range weather forecasting and air quality modeling); computational biology (genome sequence assembly, microarray data analysis, structure-based drug discovery, protein folding and molecular dynamics simulation); computational fluid dynamics (viscous, compressible, unsteady flows; laminar natural convection flows; fluid flow and heat transfer for heat exchangers and simulation of fire in high rise buildings) and computational structural mechanics (stress analysis of fibre reinforced composite structures, fracture mechanics, nonlinear stability analysis, seismic vulnerability analysis, hazard assessment of civil structures).

Garuda is India's national grid infrastructure of HPC systems, connecting 70 academic and research institutions across 17 cities of the country with India's Nation Knowledge Network (NKN). The Indian Grid Certification Authority (IGCA) accredited by the APGrid-PMA (Asia Pacific Grid Policy Management) has been set up in C-DAC. The IGCA helps scientists, researchers and collaborative community in India and neighboring countries to obtain an internationally recognized digital certificate to interoperate with state of the art grids worldwide. Some of the applications that make use of this grid infrastructure include: Open Source Drug Discovery (OSDD); collaborative class room; computer aided engineering; oncology research; disaster management using synthetic aperture radar; biodiversity conservation and cloud computing.

“Cloud Computing would be a solution for big data problem”, was opined by the experts at the Indian Science Congress held at Mysore . C-DAC's research focus in cloud computing includes design and development of open source cloud middleware; virtualisation and management tools; and end to end security solution for the cloud. A number of applications in C-DAC are being migrated to cloud computing technology. These include hospital information systems, disaster recovery, telemedicine, HPC services, language services, e-governance applications and many more. Considering the related, but complimentary driving forces of grid and cloud computing disciplines, C-DAC is also exploring integration of grid and cloud computing.

With the expertise of C-DAC on various technologies that act as enablers of cloud computing and National Resource Centre For Free and Open Source Software (NRCFOSS), C-DAC has developed a comprehensive free and open source suite named 'Meghdoot' for setting up a cloud computing environment.

On analysing various free and open source tools for cloud, the featuring tools across all layers of

the cloud were incorporated into the suite. C-DAC pursued the remarkable in-house development of features and functionalities based on standards over free and open source tools of cloud. Research efforts have improved the existing open source cloud tools. The developments include, service provisioning and deployment, management, security and other value additions. Simple graphical installation and configuration of the suite, exhaustive monitoring, metering, customisable elasticity, web service oriented management and inclusion of security features focusing on data in transit, data at rest, multi level authentication and authorisation are a few of the notable value additions by C-DAC.

The first version of the suite was released at the C-DAC technology conclave, Hyderabad in October 2012. The product has been well received by users from all sectors - banking, e-governance, academic and research, telecommunications, healthcare, logistics based SME, corporate, and manufacturing.

Indian Banking Community Cloud, established by Institute for Development and Research in Banking Technology (IDRBT), Hyderabad operates on Meghdoot. This community cloud incorporates various customised developments specific to the banking community requirements. It was inaugurated by the Governor, RBI, and is operational since August 2013, supported by C-DAC. Up till recently, about 13 banks have hosted their applications in this cloud.

In June 2016, Comptroller and Auditor General who also overlooks various financial aspects of the Indian Armed Forces announced setting up a data analytics center which is first of its kind by any auditor, for analysing big data in the government domain.

To impart knowledge to the aspirants, various training programmes are conducted by C-DAC. Customized corporate training programme was also conducted for Indian and foreign organisations. C-DAC conducts various training workshops in collaboration with academic institutions in the country. C-DAC also conducts lectures, seminars, pre-conference tutorials in national and international conferences.

Likely Applications for the Defence and Security Forces Organisations

A pioneering work titled “Big Data–Applicability in the Defence Forces” was published as a seminar report by the Centre for Land Warfare Studies (CLAWS), New Delhi in March 2015 . Sectors Identified for implementation of big data applications in defence forces were: intelligence and surveillance; border, maritime and space management; operational planning; logistics management; fiscal and financial management; disaster management; future technologies; cognitive analytics and analysis of archival data. Suggested future scope of work was: creation of centre of excellence for big data initiatives with a Chair, setting up protocol for sharing of simulated data and signing non disclosure agreements, organise round tables with defence forces for development of visualisation tools, big data training and policy formulation. A perceived

roadmap suggesting implementations in six stages addressing all requirements of the Indian Army has also been given. These applications cover almost all important functions of the defence forces which are presently being addressed by several data base management tools available in the market. It has not yet been established whether defence forces have the capacity and capability to generate big data (terabyte to pent byte range) on a sustained basis, necessitating employment of BDA and creation or sharing of BDA infrastructure.

However, fact remains that due to nascent nature of big data analytics its awareness is limited to a small number of involved agencies. The benefits of big data in operational scenario decision making while safeguarding accuracy and reliability have not yet been internalized. Big data projects even at pilot scales may not be available currently. In the present situation, decision makers are not clear about the capability of big data, costs, benefits, applicability or the perils if any of not adopting big data .

Traditional approach of slow and structured employment of emerging technologies is unlikely to benefit the defence forces as software applications are likely to outpace the traditional approach. As an important step, the C-DAC should be persuaded to conduct intensive on site short training capsules for the senior decision makers in the MoD and defence HQs. Simultaneously, ICT set up officers and personnel should be trained at C-DAC for understanding the employability and intricacies of the BDA related infrastructure.

Many organisations are already making use of the C-DAC, BDA related infrastructure. C-DAC is open to participation by more organisations including defence forces. Reasonable security features are incorporated in the system. Banks have organised themselves in a big way for using the C-DAC infrastructure. Defence forces could follow their example.

BDA has the potential to assist defence forces in almost all important functions as suggested by the CLAWS paper also. Understanding the capability of the BDA especially at senior level is essential for effective utilisation of its applications. Training and more training is needed for this purpose. In addition to C-DAC, a number of private universities are offering online training courses on BDA. A number of private institutions are also offering training on various soft wares such as Hadoop. This opportunity can be availed by the far sighted commanders and ICT specialists in the defence forces as well.

Technological developments have always added new dimensions to the war waging capabilities. Fast changing developments in the ICT field have shortened the OODA loop and added transparency to the battle space. Like science and technology disciplines, data proliferation in the defence and security forces has also multiplied many times. Ever increasing sensors and networking on high speed broadband networks coupled with fast data processing capabilities has begun to demystify the fog of war and enhance weapon system effectiveness.

The so called data deluge has begun to cross the boundaries of traditional data base management systems. Necessity of BDA has already begun to be felt especially in the ISR field. US, NSA and other defence organisations are already using the BDA in a big way.

The emerging BDA eco system in India is presently centered on the DST and C-DAC. Various other departments and organisations are using this infrastructure to meet their requirements. The C-DAC infrastructure consisting of super computers, high speed networks and secure cloud system offers an interesting opportunity for the defence and security forces to be benefitted from the BDA applications.

Big Data Analytics Training

There are several emerging technologies which are finding applications in the BDA. Department of Science and Technologies (DST) is attempting to ensure that necessary awareness is created among potential stake holders. One such DST sponsored 10 days training programme on BDA was organised by Institution of Engineering and Management (IEM) Kolkata last year.

C-DAC is regularly conducting Training on "Hadoop for Big Data Analytics" and "Analytics using Apache Spark". Developers, researchers, engineers, and faculty from industry, research labs, universities and colleges, who need hands-on experience on big data analytics, attend these training courses . A few defence forces personnel also undergo this training periodically. Four-day programme enables the data analysts and data scientists to get the insight and understanding of the Hadoop and Spark platform through well designed sessions. Topics covered for "Hadoop for Big Data Analytics" generally are: Setting up your Hadoop cluster, data management using HDFS, Map Reduce framework with Map Reduce application hands on, Apache Hive and Pig to leverage data manipulation, Apache HBASE, NoSQL Database and hands-on Apache HBASE and Apache Phoenix. Topics covered during "Analytics using Apache Spark" are; Introduction to Apache Spark, Spark architecture and internals, Large-scale parallel and distributed data processing using Spark, Spark MLlib - A Machine learning library, Spark streaming and Analytics using Spark - Use cases. C-DAC also offers short term courses of one day to one month on topics such as Database Security, Ethical Hacking, Perimeter Security, Security Engineering, Web Application Security, Wireless Security, Security Administration Linux, Cyber Forensics, Cyber Crime, IT Law, Mobile Security etc.

Privacy and Security Concerns

In an article titled “More Info, More Problems: Privacy and Security Issues in the Age of Big Data”, Jason Parms states that emerging big data scenarios have caused privacy and security concerns. He has suggested several precautions which would address many privacy and security concerns. Maintenance of security in distributed computing frameworks, providing best security practices for non relational data stores, preserving the privacy in data mining and analytics,

providing encrypted data centric security, ensuring granular access control, securing storage and transaction logging, facilitating data provenance and verification, provision for endpoint input validation and filtering and real time security monitoring are some of the measures for addressing privacy and security concerns.