Review Article

Military Simulation Big Data: Background, State of the Art, and Challenges

Xiao Song, Yulin Wu, Yaofei Ma, Yong Cui, and Guanghong Gong

School of Automation Science and Electrical Engineering, Beihang University, No. 37 Xue Yuan Road, Hai Dian District, Beijing 100191, China

Correspondence should be addressed to Xiao Song; songxiao@buaa.edu.cn

Big data technology has undergone rapid development and attained great success in the business field. Military simulation (MS) is another application domain producing massive datasets created by high-resolution models and large-scale simulations. It is used to study complicated problems such as weapon systems acquisition, combat analysis, and military training. This paper firstly reviewed several large-scale military simulations producing big data (MS big data) for a variety of usages and summarized the main characteristics of result data. Then we looked at the technical details involving the generation, collection, processing, and analysis of MS big data. Two frameworks were also surveyed to trace the development of the underlying software platform. Finally, we identified some key challenges and proposed a framework as a basis for future work. This framework considered both the simulation and big data management at the same time based on layered and service oriented architectures. The objective of this review is to help interested researchers learn the key points of MS big data and provide references for tackling the big data problem and performing further research.

1. Introduction

Big data technology is an emerging information technology which discovers knowledge from large amounts of data to support decision-making. "Big" implies that the resulting dataset is too large to be handled with traditional methods. Moreover, the data production velocity is typically fast with various sources [1]. Big data methodology is even regarded as the fourth paradigm of exploring the world, since it is different from experimentation, theoretical approaches, and computational science in terms of knowledge acquisition methods [2].

Recently, big data and simulation have been linked by researchers to perform scientific discovery [3, 4]. Military domain has also drawn a great deal of attention to this trend. The United States (US) Department of Defense (DoD) is carrying out a series of big data programs (e.g., XDATA) to enhance defense capabilities [5]. Generally, military applications are producing massive amounts of data with plenty of Intelligence, Surveillance, and Reconnaissance (ISR) sensors [6, 7], and the data can also be generated by Live, Virtual,

and Constructive simulations [8]. Besides, all kinds of data about combat entities and events in battlefield are collected together.

Models and simulations (M&S) are often classified by US DoD into four levels: campaign (theater), mission, engagement, and engineering, usually depicted as a "M&S pyramid" [9, 10]. M&S applications span all levels of this pyramid, with campaign models and simulations being applied in warfare analysis, mission-level simulations in such areas as Joint Operations for analysis or exercise [11–15], engagement simulations in confrontation of system of systems for warrior training, and engineering-level models and simulations in equipment acquirement and development for test and evaluation (T&E).

When these manifold models and simulations are executed with high performance computing (HPC) to gain high efficiency, simulation data has the great potential to be generated with high volume and rapid speed. Recently, the term "big simulation" was coined by Taylor et al. [16] to describe the challenge that models, data, and systems are so excessive in scale that it is difficult to govern the simulation.

Accordingly, we regard military simulation data as "*MS big data*" to describe the high volume of data generated by military simulations.

MS big data can be produced by numerous simulation replications containing high-resolution models or large-scale battle spaces or both. The data size increases rapidly with larger simulation and higher performance computer resources. This poses significant challenges in management and processing of MS big data. First, great efforts have been made to address the requirements from high performance simulation, while only a few is toward data processing. However, the processing of MS big data makes some differences between business data; for example, the computing resources are often used for data analysis together with simulation execution and this may lead to more complex resource scheduling. Second, new requirements will emerge with availability of high fidelity models and timeliness of large amounts of data. Traditional data analytic methods are limited by traditional database technologies with regard to efficiency and scalability. Now big data based new applications allow military analysts and decision-makers to develop insights or understand the landscape for complex military scenarios instead of being limited to only examining experimental results. An example is supporting the real-time planning of real combat by simulating all kinds of possibilities. This requires iterative simulation and analysis of result in very short time.

This paper serves as an introduction to the leading edge of MS big data. There are already many review papers discussing the concept of big data and its underlying technical frameworks [17–19]. However, few of them focus on military simulation, yet related researchers should be interested in the generation, management, and application of MS big data. We will study some practices and development progress by literature so that a general picture of MS big data can be drawn. Meanwhile, for unique characteristics of MS big data, we will identify some technique challenges posed by the processing of MS big data. We will also demonstrate a preliminary framework as possible solution.

The remainder of this paper is organized as follows. The next section introduces the background of MS big data. Several practical cases are reviewed, along with a summary of the characteristics of MS big data. In Section 3, the detailed advancements in technology are discussed. In Section 4, we investigate two platforms closely related to MS big data. Current challenges are identified in Section 5, along with a proposed framework to integrate simulation and big data processing. Finally, we conclude the paper in Section 6.

2. Background

This section surveys the background that MS big data problem emerged, but firstly we must address what big data is in general. Then we review several military simulations from viewpoint of data. Finally the features of MS big data are discussed and compared with those in business.

2.1. Concept of Big Data. Big data refers to the data set which is so huge that new processing methods are required to enable knowledge discovery. The term implies the trend of data explosion and its potential value for present society. Data scientists often use N-V (volume, velocity, variety, value, veracity, etc.) dimensions to account for big data.

Volume. The size of big data is not exactly defined. It varies from different fields and expands over time. Many contemporary Internet companies can generate terabytes of new data every day and the active database in use may exceed petabyte. For example, Facebook stored 250 billion photos and accessed more than 30 petabytes of user data each day [21]; Alibaba, the biggest electronic commerce company in the world, had stored over 100 petabytes of business data [22]. Large volume is the basic feature of big data. According to International Data Corporation (IDC), the volume of digital data in the world reaches zetabytes (1 zetabyte = 2^{30} terabytes) in 2013; furthermore, it almost doubles every two years before 2020 [23].

Velocity. Data explosion also means that the data generation speed is very quick and they must be processed timely. Take Facebook, for example, again, millions of content items are shared and hundreds of thousands of photos are uploaded every minute of the day. To serve one billion global users, over 100 terabytes of data are scanned every 30 minutes [24]. Velocity is also a relative concept and depends on practical application. For many Internet commerce applications, the processed information must be available in a few seconds; otherwise the value of data will diminish or be lost.

Variety. Big data has diverse types and formats. Tradition structured data saved in database possess regular format, that is, date, time, amount, and string, while unstructured data such as text, audio, video, and image are main styles of big data. The unstructured data can be web page, blog, photo, comment on commodity, questionnaire, application log, sensor data, and so forth. These data express human-readable contents but are not understandable for machine. Usually, they are saved in file system or NoSQL (Not Only SQL) database which has simple data model such as Key-Value Pair. Variety also means that big data have various sources. For example, the data for traffic analysis may come from fixed network camera, bus, taxi, or subway sensors.

Value. Big data can produce valuable insight for owner. The insight help to predict the future, create new chance, and reduce risk or cost. As a result, big data can change or improve people's life. A famous example of mining big data is that Google successfully forecasted flu according to the 50 million search records. Another example is that the commodity recommendation can be found everywhere when we surf in Internet nowadays. These recommendation items are generated according to large amounts of records about user access. Note that the value of big data is sort of low density and must be extracted from its huge volume.

Veracity. Veracity means only that the trustworthy data should be used; otherwise the decision maker may get false knowledge and make wrong decision. For example, the online

review from customer is important to ranking system of commodity, and if some people make fake comments or score deceptively for profit, the result will influence negatively the ranking and customer's choice. Veracity requires those fake data to be detected and eliminated prior to analysis.

Big data emerged from not only Internet social media or commerce but also government, retailing, scientific research, national defense, and other domains. Nowadays they are all involved in processing of massive data. The bloom of big data application is driving the development of new information technology, such as data processing infrastructure, data management tool, and data analysis method. Recent trend in those enterprises and organizations owning large dataset in data center is becoming a core part of information architecture, on which scalable and high performance data processing framework is running. Many of these frameworks are based on Apache Hadoop ecosystem, which uses MapReduce parallel programming paradigm. At the same time, many new data mining and machine learning algorithms and applications are proposed to make better knowledge discovery.

2.2. MS Big Data Cases.

The phenomenon producing massive data in military simulation can be traced back to the 1990s when STOW97 was a distributed interactive simulation for military exercises. It incorporated hundreds of computers and produced 1.5 TB of data after running 144 hours [25]. As simulation technology advances, many military applications are producing multiple terabytes of data or more.

2.2.1. Joint Forces Experiments.

When DoD realized the challenges of contemporary urban operation, the US Joint Forces Command commissioned a series of large-scale combat simulations to develop tactics for the future battlefield and evaluate new systems (e.g., ISR sensors) deployed in an urban environment. A typical experiment was Urban Resolve with three phases across several years [11]. It used the Joint Semiautomated Forces (JSAF) system to study the joint urban operation, set in the 2015–2020 timeframe. JSAF is a type of Computer Generated Forces (CGF) software which generates and controls virtual interactive entities by computers.

In the first phase, more than 100,000 entities (most were civilian) were simulated. Hundreds of nodes running JSAF were connected across geographical sites, and 3.7 TB of data was collected from models [12]. These data are relevant to the dynamic environment, operational entities (including live and constructive), and sensor outputs [26]. The analysis involved query and visualization for a large data set, such as the Killer/Victim Scoreboard, Entity Lifecycle Summary, and Track Perception Matrix.

However, more data about civilian status which were saved dozens of times were simply discarded because of unaffordable resources. Therefore, civilian activity could not be duplicated for analysis. Meanwhile, the demands for larger and more sophisticated simulations were increasing. With more power clusters, graphics processing unit (GPU) acceleration, and high-speed wide area networks (WANs) using interest-managed routers [13], tens of millions of entities and higher-resolution military models were supported in later

experiments. In this case, new data management tool based on grid computing technology was proposed to address the problem from data increase [14].

2.2.2. Data Farming Projects.

Data farming is a process using numerous simulations with high performance computing to generate landscapes of potential outcomes and gain insight from trends or anomalies [15]. The basic idea is to plant data in the simulation through various inputs and then harvest significant amounts of data as simulation outputs [27]. Data farming was first applied in the Albert project (1998–2006) of the US Marine Corps [28]. The project supported decision-making and focused on questions such as "what if" and "what factors are most important." These kinds of questions need holistic analysis covering all possible situations; however the traditional methods are unable to address them because one simulation provides only a singular result [20]. By contrast, data farming allows for understanding the entire landscape by simulating numerous possibilities.

The number of simulation instances running concurrently through HPC usually is large. For example, a Force Protection simulation of German Bundeswehr created 241,920 replicates [29]. Furthermore, millions of simulation runs can be supported by the latest data farming platform built on heterogeneous resources including cluster and cloud [30]. As a result, massive data could be produced for analysis [20]. For example, the Albert project generated hundreds of millions of data points in its middle stage [31].

During the Albert project, many countries in the world have leveraged the idea to study all kinds of military problems. Several simulation systems for data farming were established. Some examples are MANA of New Zealand [32], PAXSEM of Germany [33], and ABSNEC of Canada [34]. The research fields involved command and control, human factors, combat and peace support operations net-centric combat, and so forth.

2.2.3. Course of Action Analysis.

In order to effectively complete mission planning, it is crucial to recognize certain key factors of the battle space via simulations. It is especially important that the military commander evaluates possible plans and multiple decision points. This kind of experiment is called simulation-based Course of Action (COA) analysis and needs to test many situations with a large parameter value space [35]. The simulation always runs faster than real-time, and it can be injected with real-time data from actual command and control systems and ISR sensors. For high-level COA analysis, a low-resolution, large entity-count simulation is used. Furthermore, a deeper and more detailed COA analysis needs models with higher resolution. During peacetime, the COA can be planned carefully with more details. But during a crisis, the COA plan must be modified as necessary in a very short amount of time [36].

The US army uses COA analysis in operations aiming at urban environments. OneSAF (One Semiautomated Forces) is a simulation system which fulfills this type of requirement. As the latest CGF system of the US Army, it represents the state of the art in force modeling and simulation (FMS) [37],

including capabilities such as behavior modeling, composite modeling, multiresolution modeling, and agent-based modeling. It is able to simulate 33,000 entities at the brigade level [38] and has been ported to HPC systems to scale up with higher resolution models [39]. Real system can be integrated with OneSAF through adapters so that the simulation is enhanced. Massive data can be generated to analyze and compare different COA plans. In this case, data collection and analysis are identified as its core capabilities [40].

The Synthetic Theater Operations Research Model (STORM) is also an analysis tool applicable to COA. It simulates campaign level and is used by US naval force and air force. It can create gigabytes of output from single replication, and one simulation experiment may contain a set of replications running from several minutes to hours, depending on the complexity of scenario and models [41, 42].

2.2.4. Acquisition of New Military Systems. Australia's acquisition of naval systems [43, 44] employed modeling and simulation (M&S) to forecast the warfare capabilities of antiair, antisurface, and antisubmarine systems since the real experiment costs high or was unavailable. The M&S played a significant role in the acquisition lifecycle of new platforms, such as missiles, radar, and illuminators. All phases, including requirements definition, system design and development, and test and evaluation, were supported by M&S using computer technology. The simulation software contained highly detailed models, which can be used for both constructive and virtual simulation at a tactical level. The simulation scenarios were defined by threat characteristics, threat levels, and environmental conditions.

The project spanned across several years, involving multiple phases and stages. It also involved multiple organizations across multiple sites. A large-scale simulation with hundreds of scenarios (each scenario ran hundreds of times with Monte Carlo simulation) was executed on IBM blade servers and terabytes of data were generated. Complex requirements for analyses of the large dataset, such as verifying assumptions, discovering patterns, identifying key parameters, and explaining anomalies, were put forward by subject matter experts. These works generated Measures of Effectiveness (MOE) to define the capabilities of the new systems.

2.2.5. Space Surveillance Network Analysis. The SSNAM (Space Surveillance Network and Analysis Model) project was sponsored by the US Air Force Space Command, and it used simulations to study the performance and characteristics of the Space Surveillance Network (SSN) [45]. The purpose was analyzing and architecting the structure of the SSN. A number of configuration options, such as operation time, track capacity, and weather condition, were available for all modeled sensors in the SSN.

One research problem was Catalog Maintenance. SSNAM provided capabilities to assess the impacts from catalog growth and SSN changes related to configuration and sensors (e.g., addition, deletion, and upgrade). This kind of simulation was both computationally and data intensive. A typical run simulated several thousands of satellites within a few

hours in terms of wall-clock time. The simulation was super-real-time, and the simulated time itself could be 90 days. The original simulation results data reached the terabyte level. Its analysis was measure of performance (MOP) based on recognized parameters from daily operations. The system was a networked program based on a load-sharing architecture which was scalable and could include heterogeneous computational resources. To reduce the execution time, SSNAM has been ported to an HPC system and gained three times increase in performance as a result.

2.2.6. Test and Evaluation of the Terminal High Altitude Area Defense System. The mission of the Terminal High Altitude Area Defense (THAAD) system was to protect the US homeland and military forces from short-range and medium-range ballistic missiles. The test and evaluation of the THAAD system were challenged with analyzing the exponentially expanding data collected from missile defense flight tests [46]. The system contained a number of components (e.g., radars, launchers, and interceptors) and was highly complex and software-intensive. Two phases have been evaluated by experiment: deployment and engagement. The THAAD program incorporated the simulation approach to support system level integrated testing and evaluation in real-time. In this case, simulation was used to generate threat scenarios, including targets, missiles, and environmental effects. In addition, M&S was also used for normal exercises.

The THAAD staff developed a Data Handling Plan to reduce, process, and analyze large amounts of data. ATHENA software was designed to manage the terabytes of data generated by flight tests and concomitant simulations. The ATHENA engine imported various files such as binary, comma-separated values, XML, images, and video format. The data sources could come from LAN, WAN, or across the Internet.

2.3. Characteristics and Research Issues of MS Big Data. Table 1 summarizes the main features of MS big data within the above cases.

We can draw the main characteristics of MS big data from the above cases in terms of volume, velocity, variety, and veracity.

Volume and Velocity. Table 1 shows the data sizes. Almost all cases are at the level of GB to TB per experiment (see Table 1(b)). These simulation cases' data volume is smaller compared with commerce and social media on internet and web, because simulation experiment can be well designed and the needed data for analysis can be chosen carefully to save. It suggests that the value density of simulation data is higher. Another reason is that simulation data are not accumulated across experiments; different experiments have different objectives and are seldom connected to do data analysis.

However, the simulation data size continues to increase sharply because military simulation is advancing rapidly with bigger scale and higher resolution under the impetus from modern HPC system. Moreover, most business data are produced in real-time [47], but MS big data needs to

TABLE 1: Overview of military big simulations.

(a) Simulation-related properties

Cases	Simulation level				The number of entities	The number of simulation replications	Model resolution	Time advancing	Simulation execution time
	C	M	E	G					
Joint forces experiments	✓	✓	✓		Millions	Several	High	Real-time	Weeks
Data farming projects	✓	✓	✓		Dozens~hundreds	Thousands~millions	Low	Faster-than-real-time	Hours
COA analysis	✓	✓	✓		Thousands	Hundreds~thousands	Low medium	Faster-than-real-time	Minutes hours
Acquisition of new military system			✓	✓	A few	Tens of thousands	High	Faster-than-real-time or real-time	Days or weeks
SSN analysis				✓	Tens of thousands	Several	High	Faster-than-real-time	Days
Test and Evaluation of the THAAD system			✓	✓	A few	Several	High	Real-time	Months

Simulation level: C, campaign; M, mission; E, engagement; G, engineering.

(b) Data-related properties

Cases	Data contents	Data generation period	Data scale per experiment	Typical analytical application
Joint forces experiments	Sensor data, entity status (take Urban Resolve, e.g.)	1 ms	TB	Effectiveness of ISR sensor. Statistical result of mission execution
Data farming projects	Depend on specific problem and can involve all kinds of physical and behavioral data of simulated entities and battlefield events, such as damage and survival	<1 ms	GB~TB	Correlation, trend, and outlier analysis related to command and control, peacekeeping operation, combat, and so forth
COA analysis	Civil behavior, culture, weapon, terrain, killing, victims, and so forth	<1 sec	GB~TB	Task execution analysis, operational effectiveness analysis
Acquisition of new military system	Communication, sensor, command and control, weapon, and so forth	N/A	TB	Identify the key performance parameter, system level trade-off analysis, and MOE about raid, protect, detect, reaction, tracking, engagement, and so forth
SSN analysis	Sensor data and special events such as launch, breakup, and in-orbit	<1 sec	TB	MOP of sensor device
Test and evaluation of the THAAD system	Maneuver, sustainment, command and control, communication, radar track data, and so forth	N/A	TB	MOE, MOP of defense system

be generated and analyzed many times faster than real-time when the objective is to rapidly assess a situation and enhance decision-making. This requires simulation time to advance faster than real-time (see Table 1(a)), and sometimes the simulation generates data in a period of less than 1 ms (see Table 1(b)).

High volume and velocity pose two aspects of challenges to MS big data applications. First, collecting massive data from distributed large-scale simulations may consume extra resources in terms of processor or network, which is often critical for simulation performance. Thus it is essential to design a high-speed data collection framework that has little impact on the simulation performance. Second, the datasets must be analyzed at a rate that matches the speed of data production. For time-sensitive applications, such as situation awareness or command and control, big data is injected into the simulation analysis system in the form of a stream, which requires the system to process the data stream as quickly as possible to maximize its value.

Variety. Large-scale simulations can be built based on the theory of system of systems (SoS) [48], which consists of manifold system models such as planes, tanks, ships, missiles, and radars. Figure 1 presents various kinds of data involved in military experiments.

From Figure 1, we can observe the diversity of MS data. For example, MS can be linked with live people (e.g., human-in-the-loop) and live military systems (e.g., command and
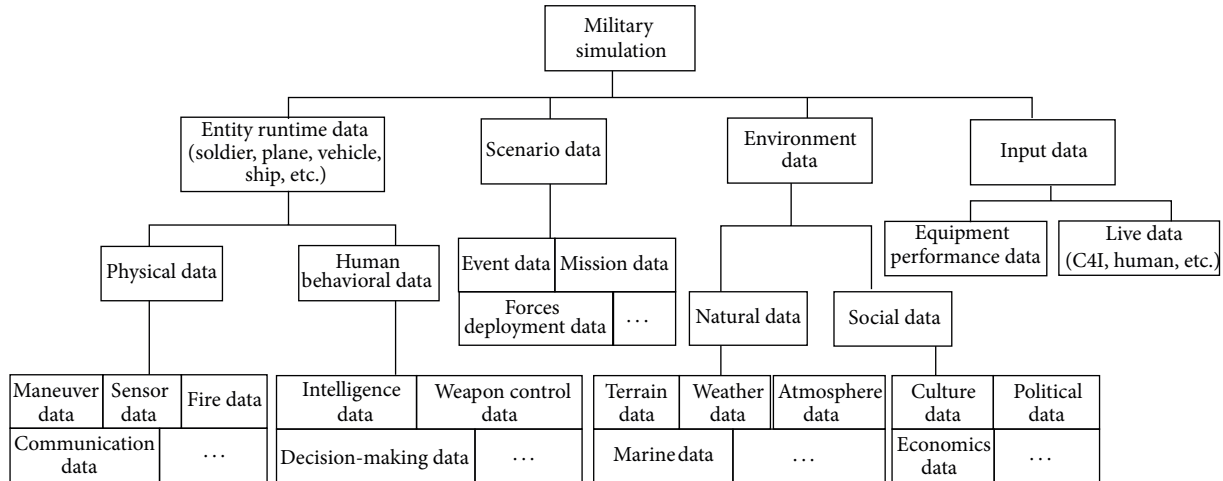
FIGURE 1: Various kinds of data in simulation-based military experiment.

control devices). As such, valuable analytic MS data includes outputs of computational models, as well as human activity and device data. The data formats include unstructured (e.g., simulation log file), semistructured (e.g., scenario configuration and simulation input), and structured (e.g., database table) types. All of these features require versatile and flexible tools to be developed to mine value from the data effectively. This imposes difficulty on the data processing technology.

Veracity. Veracity means that trustworthy data should be created during the simulation. Because simulation data is generated by computer but not human, the data could not be fake but can be incorrect because of flawed model. Veracity requires that the model and input data should be verified and validated. However, the difficulty is that MS big data is often involved in human behavior, which is intelligent, yet intangible and diverse by nature. Compared with analytic physical or chemical models, there is no proven formula that can be used for behavior modeling. Currently a behavior model can create data only according to limited rules recognized by humans. Therefore, the fidelity of simulation models is a key challenge for the veracity of MS big data.

The next section will give an overview of the MS big data technology, where we can see how the above problems are resolved to some extent.

3. State of the Art of MS Big Data Technology

From the viewpoint of data lifecycle, we can divide the simulation process into three consecutive phases: data generation, data management, and data analysis. Figure 2 shows the detailed technology map. Data generation concerns what kinds of data should be created and how to create valid data in a reasonable amount of time. Data management involves how to collect large amounts of data without disturbing the normal simulation and provide available storage and efficient processing capability. Data analysis utilizes various analytic methods to extract value from the simulation result.
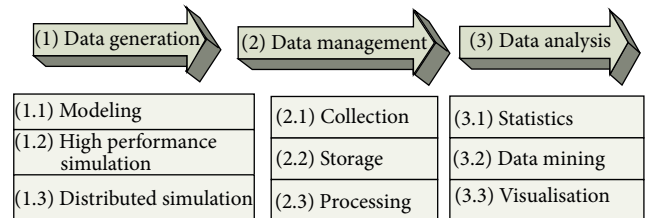

FIGURE 2: Technology map of MS big data.

3.1. Data Generation

3.1.1. Combat Modeling. Modeling is a key factor to the veracity of MS big data [28]. In combat modeling, the main aspects are physical, behavioral, environmental, and so forth, and the behavior model is the most sophisticated part.

Many research efforts focus on developing a cognitive model of humans, a Human Behavior Representation (HBR), which greatly affects the fidelity and credibility of a military simulation. HBR covers situation awareness, reasoning, learning, and so forth. Figure 3 shows the general process of human behavior. Recent research projects include the following: situation awareness of the battlefield [49]; decision-making based on fuzzy rules, which captured the approximate and qualitative aspects of the human reasoning process [50]; common inference engine for behavior modeling [51]; intelligent behavior based on cognition and machine learning [52]; modeling cultural aspects in urban operations [53]; modeling surprise, which affects decision-making capabilities [54, 55]. Although it is still difficult to exhibit realistic human behaviors, the fidelity of these models can be enhanced by adoption of big data analytic technologies.

Meantime, some researchers focus on the emergent behaviors of forces as a whole. A battlefield is covered by fog due to its nature of nonlinearity, adaptation, and coevolution [56]. Agent-based modeling (ABM) is regarded as a promised technique to study such complexity [57] by simulating autonomous individuals and their interactions so that
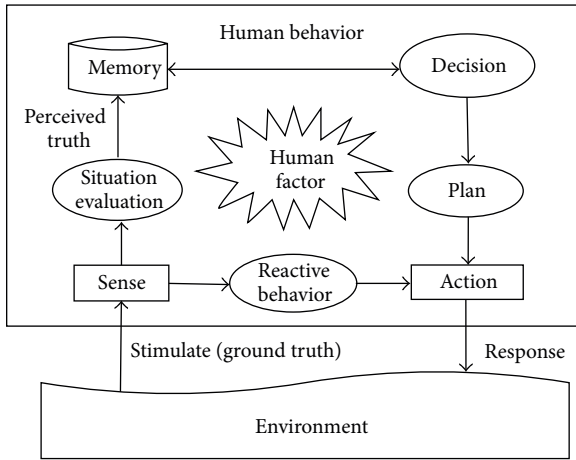
FIGURE 3: The procedure of human behavior.

the whole system can be evaluated. The typical interactive behaviors are command and control (C2), cooperation and coordination, and offensive and resistance. Recent research efforts include logical agent architecture [58], artificial intelligence based on agent [59], and multiagent framework for war game [57, 60]. An interesting point is that agent-based systems are considered to have strong connections with big data [16], because ABM provides a bottom-up approach in which modeling large amounts of individuals and their emergent behavior must be identified from the big data generated.

3.1.2. High Performance Simulation. To address the enormous computational requirements of large-scale military simulations, a high-performance computing (HPC) technique is employed as a fundamental infrastructure [61–65]. For example, DoD has developed the Maui High Performance Computing Center (MHPCC) system, which includes large-scale computing and storage resources, to support military simulations. In another typical example, the three-level parallel execution model (Figure 4) was proposed [61]. In Figure 4, the top level is parallel execution of multiple simulation applications, the middle level is parallel execution of different entities within the same simulation application, and the bottom level is parallel execution of different models inside one entity.

In spite of its huge processing power, the cluster-based HPC systems still remain considerably challenged to handle the data-intensive issues [66] presented by simulation applications. To address this issue, Hu et al. [67] proposed developing advanced software specific to simulation requirements because there is a lack of software and algorithms to handle large-scale simulations. Also, many in the field think that simulation management software should provide functions such as job submission, task deployment, run-time monitoring, job scheduling, and load balancing.

Some approaches have begun to address this problem. For instance, a new HPC system named YH-SUPE has been implemented to support simulation [61] by optimizing both hardware and software. The hardware contains special components used to speed up specific simulation algorithms,

and the software provides advanced capabilities, such as time synchronization based on an extra collaborative network and efficient scheduling based on discrete events. Another approach is to port an existing simulation system onto a common HPC platform. OneSAF had a good experience with this approach [62–64]. In addition, enhancing a model's performance (e.g., at performing line-of-sight calculations) using a GPU accelerator is also a research hot spot [65].

3.1.3. Distributed Simulation. Distributed simulation often uses middleware to interconnect various simulation resources, including constructive, virtual, and live systems. Middleware means that the tier between hardware and software usually is an implementation of a simulation standard for interoperation. For historical reasons, several standards are now being used: distributed interactive simulation (DIS), high-level architecture (HLA), and Test and Training Enabling Architecture (TENA). HLA is an upgraded standard for distributed simulation comparable to DIS. TENA is used to test military systems and personnel training, but HLA is not limited to the military domain.

Middleware enables large-scale simulations to execute with a large number of entities from different nodes, such as HPC resources. Middleware is the key to scalability of the simulation, and many researchers have focused on the performance improvements it makes possible. For example, enhancement work on the Runtime Infrastructure (RTI, the software implementation of HLA) with regard to Quality of Service and data distribution management is very interesting because it makes proper use of HPC resources [68].

However, the interconnected applications may introduce geographically distributed data sources. This also poses the complexity of managing and exploiting MS data for large regions. To tackle this problem, [14] proposed a two-level data model: the original collected data were organized by the logging data model (LDM) and then transferred into the analysis data model (ADM). ADM represented the notion of an analyst and was defined as Measures of Performance (e.g., sensor effectiveness), which involved multiple dimensions of interests (e.g., sensor type, target type, and detection status). Another approach is to establish a standard format which covers all kinds of data in specific applications [69, 70].

3.2. Data Management

3.2.1. Data Collection. When a large-scale simulation is executed by an HPC system, the compute-intensive models can generate a deluge of data to update entity state and environment conditions [71]. The transfer and collection of data generate data-intensive issues. As the data are created by simulation programs, the collection is executing as simulation logging. This logging is undergoing an evolutionary change from a standalone process to a fully distributed architecture (Figure 5).

In the standalone logger method, one or several nodes are established to receive and record the published data via the network [72]. However, this approach aggravates the shortage of network resources because it requires large amounts of
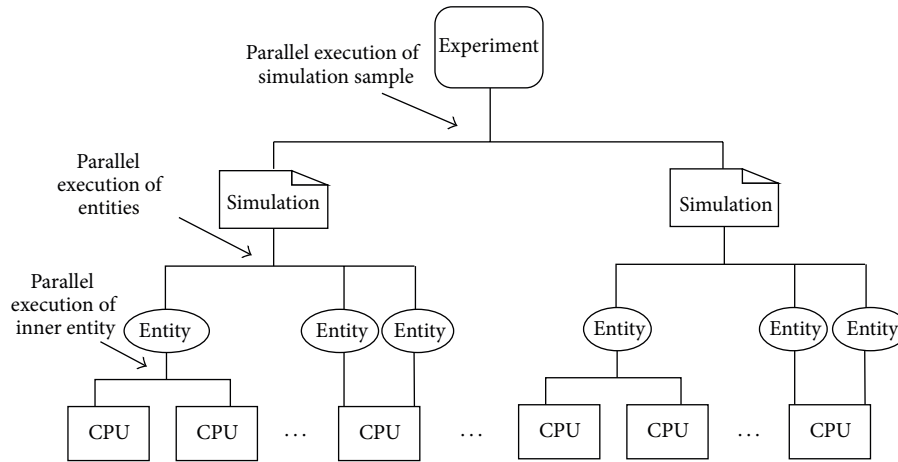
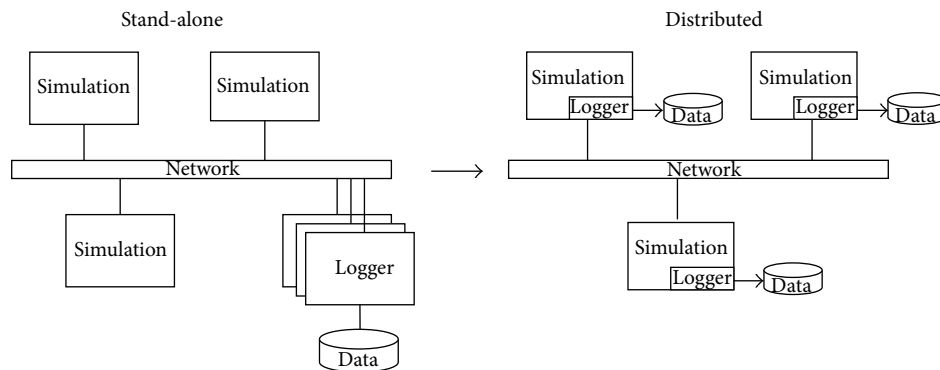FIGURE 4: Three-level parallel execution of simulation experiment.

FIGURE 5: Change from stand-alone to distributed logging architecture.

data to be transferred. When the logger is moved into the HPC platform, the network communication bottleneck is alleviated by using inter-memory communication or a high-speed network [39]. Furthermore, many useful data for analysis (e.g., inner status of entities) are not transmitted by the network so that they cannot be collected using the standalone logger method, but they can be collected when the logger is integral to the HPC.

By contrast, distributed loggers residing in most simulation nodes are preferred since data are recorded in local node and thus network resource is saved [12]. Davis and Lucas [73] pointed out the principle for collecting massive data: minimize the network overhead by transferring only required information (e.g., results of processing) and keep original data in the local node. To effectively organize the dispersed data, a distributed data manager can be employed, and we will introduce this concept and related works in the next subsection.

When standalone architecture turns into distributed data collection, the technical focus also shifts from the network to the local node. The simulation execution should not be disturbed in terms of function and performance by this architectural change. Wagenbreth et al. [14] used a transparent component to intercept simulation data from standard RTI

calls, and this way ensured the independence of data collection. Wu and Gong [74] proposed a recording technique with double buffering and scheduled disk operations based on fuzzy inference, so that the overhead of data collection can be significantly reduced.

3.2.2. Data Storage and Processing. Traditional databases have limitations in performance and scalability when managing massive data. Now that simulation scale keeps increasing, the data management must be able to easily address the requirements for future datasets. Distributed computing technology could utilize a lot of dispersed resources to provide tremendous storage capacity and extremely rapid processing at a relatively low cost point. This method is very suitable for the case where the simulation data have been recorded dispersedly, and the term "*in-situ* analysis" is used to describe this kind of data processing [75]. Distributed technology also provides better scalability, which is an important performance attribute for large-scale systems.

Currently several typical storage and processing methods can be used for MS big data.

(i) *Distributed Files.* One simple method is to utilize the original file system by saving log files in each node and then creating a distributed application which
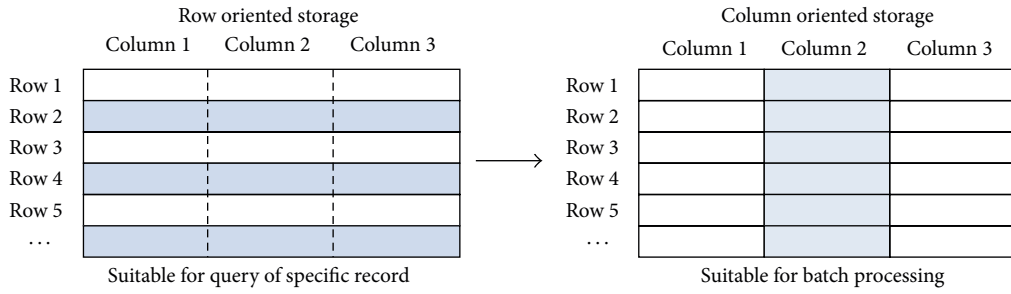
Figure 6: Row oriented database and column oriented database.

Table 2: Comparison of storage technologies for MS big data.

Storage	Example	Typical processing method	Advantages	Disadvantages
Distributed file	Specific application	Original network program	Simple	System is not generally used; Performance is not optimized
Distributed file system	HDFS	MapReduce	High-performance; scalable; reliable	Deployment is complex; processing algorithm is complex
Data grid	SDG	Grid computing	Scalable; ease of use	Complex processing is limited
NoSQL database	HBase	MapReduce-wrapped interface	Good performance; scalable; data processing is transparent	Query function is limited; system is not mature

manipulates them. In this case, both the file format and the application are specifically designed, and the scalability of data accessing is relatively low. In order to improve its universality and efficiency, a generalized list was designed to accommodate all kinds of data structures in large-scale simulations [76].

(ii) *Distributed File System.* Another method is to use a dedicated distributed file system. For example, the Hadoop distributed file system (HDFS) provides common interfaces to manage files stored in different nodes, but the user does not need to know the specific location. HDFS also provides advanced features, such as reliability and scalability. MapReduce is a data processing component compatible with HDFS. It realizes a flexible programming framework and is proved to be of high performance. The feasibility of handling massive simulation data from US Joint Forces Experiments was investigated with Hadoop [77, 78], and the result was positive, though it is pointed out that the performance of the WAN must be improved because it is a key factor of large-scale military simulations.

(iii) *Data Grid.* Database systems bring many advantages, such as simple design, powerful language, structural data format, mature theory, and a wide range of users. Although a single centralized database is not so practical for distributed simulation, multiple databases can be connected together to accommodate massive data and process them in parallel to improve the response speed. To build a cooperative mechanism among databases, data grid technology can be used

to manage tasks such as decomposing queries and combining the results. Data grids can be organized by hierarchical topology and expanded on demand. The researchers in Joint Forces Experiments constructed a data grid platform (called SDG, and the details will be discussed in Section 4) to manage distributed databases running MySQL [13, 14].

(iv) *NoSQL Database.* The NoSQL approach optimizes data accessing for data-intensive applications. There are several instances of NoSQL database systems, and we take HBase as a typical example. HBase is built based on HDFS and has column oriented style, which means that the data are not organized by row, but by column instead: values of records within the same column are stored consecutively (see Figure 6). That is because big data analysis often concerns the whole (in certain dimensions) but not a detailed record. NoSQL also optimizes data writing: simulation data do not involve complex transactions, and the records are relatively independent of each other. However, relational database systems take extra time in checking the data consistency. These unnecessary features are discarded by HBase. In addition, HBase employs a distributed architecture with load balancing capability so that data are stored and processed at flexible scale. Wu and Gong [74] presented an example which transformed the data format of a simulated entity status from database to HBase.

Table 2 compares the primary features of above methods.

3.3. Data Analysis. Data analysis includes algorithms and tools which extract information from big data and present the results to analysts. We first discuss the purpose of MS big data analysis and then review the emerging analytical methods and applications.

3.3.1. Purpose. Because of the diversity of military decision-making problems, the purpose of data analysis varies significantly. It is difficult to get a complete view of all possible subjects and related methods. Generally, the purposes can be classified into three levels according to the degree of data usage: descriptive, predictive, and prescriptive [17].

(1) Descriptive Analytics. Descriptive analytics is a primary use of a dataset and describes what has occurred. A typical case is the measure of weapon performance or system effectiveness. The analyzed data involve interactive events related to the simulated entity and its current status. Descriptive analytics can also be assisted by visualization, which presents the simulation result (e.g., scoreboard and all kinds of statistical charts) or the simulation procedure (e.g., playback with 2D or 3D situational display). Descriptive analytics usually employ traditional statistical methods when processing the original results, but the analysis process can also involve complex machine learning and data mining algorithms. For example, in virtual training the human action data need be recognized for automatic scoring.

(2) Predictive Analytics. Predictive analytics is used to project the future trend or outcome by extrapolating from historical data. For example, the casualty rate or level of ammunition consumption in a combat scenario can be predicted by multiple simulations. Two typical methods are linear regression and logistic regression. The basic idea is to set up a model based on an acquired dataset and then calculate the result for the same scenario using new data input values. Predictive analytics can also utilize data mining tools to discover the patterns hidden in massive data and then make automated classifications for new ones [79]. However, it is recognized that military problems have pervasive uncertainty, and so it is almost impossible to produce accurate predictions [28].

(3) Prescriptive Analytics. As mentioned above, accurate predictions are difficult to obtain, but military users still need to get valuable information or insights from simulations to improve their decision-making. This kind of analysis concerns all aspects of simulation scenarios. Prescriptive analytics focuses on "what if" analysis, which means the process of assessing how a change in a parameter value will affect other variables. Here are some examples: What factors are most important? Are there any outliers or counterintuitive results? Which configuration is most robust? What is the correlation between responses? These questions require deep analysis of the data and often employ data mining methods and advanced visualization tools. Usually users are inspired by the system and become involved in the exploratory process. On the other hand, a small dataset cannot reflect a hidden pattern, and only a large amount of data from multiple samples can support this kind of knowledge discovery.

3.3.2. Methods and Applications. The big data concept emphasizes the value hidden in massive data, so we focus here on the emerging technologies for data mining and advanced visualization together with their applications. Nevertheless, traditional statistical techniques have been widely applied in military simulation, and they are still useful in big data era.

(1) Data Mining. Data mining refers to the discovery of previously unknown knowledge from large amounts of data. As a well developed technology, it could be applied in military simulation to meet various kinds of analysis requirements. There are several cases:

 (1) Association rules analysis shows the correlation instead of causality among events. For example, in the context of a tank combat simulation, detecting the relationship between tank performance and operational results may be useful.

 (2) Classification analysis generates classifiers with prelabeled data and then classifies the new data by property. For example, it can be used to predict the ammunition consumption of a tank platoon during combat according to large amounts of simulation results.

 (3) Cluster analysis forms groups of data without previous labeling so that the group features can be studied. For example, it can be used to identify the destroyed enemy groups by location and type. A normal data mining algorithm may be modified to comply with the practical problem. The Albert project adopted a characteristic rule discovery algorithm to study the relationship between simulation inputs and outputs [80]. A characteristic rule is similar to an association rule, but its antecedent is predefined. Furthermore, the relevance is calculated by measures of "precision" and "recall" instead of "confidence" and "support".

In practice, the US Army Research Laboratory (ARL) used classification and a regression tree to predict the battle result based on an urban combat scenario of OneSAF. The experiment executed the simulation 228 times and defined 435 analytic parameters. Finally, the accuracy of prediction reached about 80% [81]. The Israeli Army's Battle-Laboratory studied the correlations between events generated on a simulated battlefield by analyzing the time-series, sequence, and spatial data. Various data mining techniques were explored: frequent patterns, association, classification and label prediction, cluster, and outlier analysis [82]. In addition, they proposed a process-oriented development method to effectively analyze military simulation data [83]. Yin et al. [84] mined the associated actions of pilots from air combat simulations. The key involved a truncation method based on a large dataset. As a result, interesting actions about tactical maneuvers were found. The method was also used to control flying formations of aircraft, and then a formation consistent with high quality was chosen [85]. Acay explored the Hidden Markov Models (HMM) and Dynamic Bayesian Network (DBN) technologies in the semiautomated analysis of military training data [86].
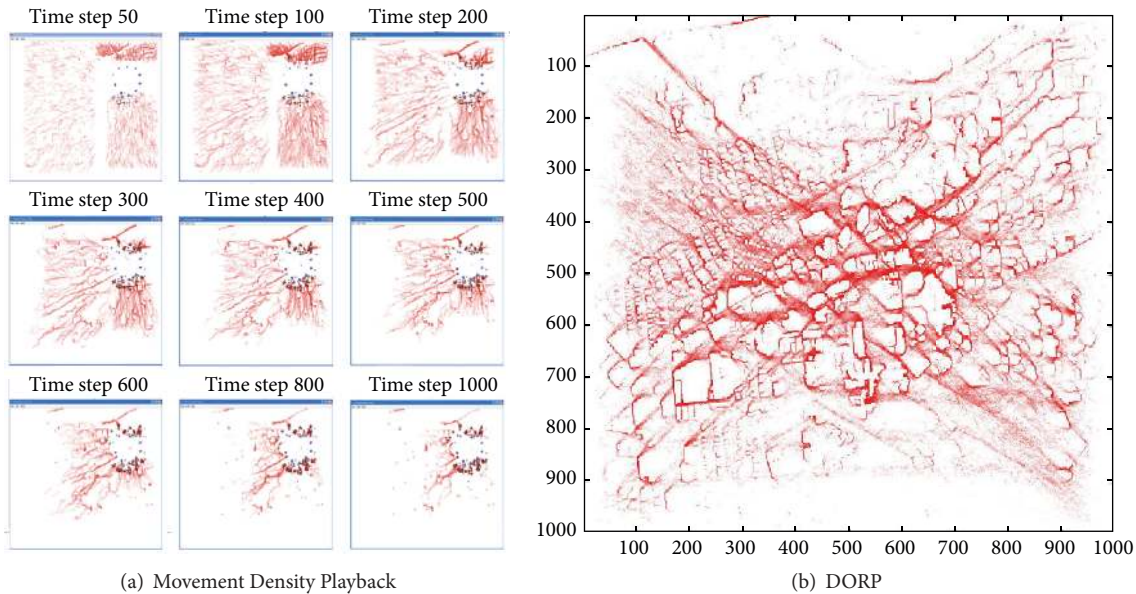
(a) Movement Density Playback

(b) DORP

FIGURE 7: Example of advanced visualization techniques (taken from [20]).

(2) Visualization. Visual analytics tools are supplements for data mining, and they are often bound together. For big data, visualization is indispensable to quickly understanding the complexity. The analyst does most of the knowledge discovery work, and the visualization tool is able to provide him with intuition and guide his analysis. All kinds of images, diagrams, and animations can be used to examine data for distribution, trends, and overall features. Furthermore, advanced visualization tools provide interactive capabilities such as linking, hypotheses, and focusing to find relevance or patterns among large numbers of parameters. In this case, the data views are dynamically changed by drilling or connecting. The research hot spots include various visual techniques built on large datasets, versatile visualization tools, and a framework for the flexible analysis of big data.

The US Marine Corps War fighting Laboratory's Project Albert [20, 87] employed different diagrams to understand simulation results: the regression tree showed the structure of the data and was able to predict the Blue Team casualties; the bar chart showed the relative importance of various combat input variables; the three-dimensional surface plots showed the overall performance measure with multiple factors; and so forth. Horne et al. [20] also reported several new presentation methods for combat simulation procedure analysis. Movement Density Playback expanded the traditional situation display by exhibiting agent trails from multiple simulation replicates (Figure 7(a)). It revealed the interesting areas/paths or critical time points of the scenario, so that the emergent behaviors or outliers could be studied. Delayed Outcome Reinforcement Plot (DORP) was a static view which showed all entities' trails from multiple replicates during their lifecycles (Figure 7(b)). It indicated the kill zone of battlefield. In addition, the Casualty Time Series chart showed the casualty count at each time step in terms of mean value from multiple simulations.

Chandrasekaran et al. [88] presented the Seeker-Filter-Viewer (S-F-V) architecture to support decision-making for COA planning. This architecture was integrated with the OneSAF system, which provided simulation data. The Viewer was regarded as the most useful component [89] because it visually tested hypotheses, such as whether an output was sensitive to specific inputs or intermediate events. An exploration environment was set up with interactive cross-linked charts so that the relationships between different dimensions could be revealed.

Clark and Hallenbeck [90] introduced the University XXI framework, which emphasized visual analysis for very large datasets. It used interactive interfaces to derive insights from massive and ambiguous data. Various data sources, data operators, processing modules, and multiview visualization were integrated and connected to check expected results and discover unexpected knowledge from operational tests, for example, ground-combat scenarios containing considerable direct-fire events.

4. MS Big Data Platforms

Based on the technologies discussed in Section 3, several simulation platforms addressing the data-intensive issues have been preliminarily developed. In this section, we will review two such pioneering platforms specially designed for MS big data: the Scalable Data Grid (SDG) manages data collected from large-scale distributed simulations, and Scalarm provides both simulation execution and data management. Each platform involves some of the technologies illustrated in Figure 2, and they use different technical architectures to structure them.

4.1. SDG. US JFCOM developed SDG to address the data problem in large simulations [13, 14]. SDG supports
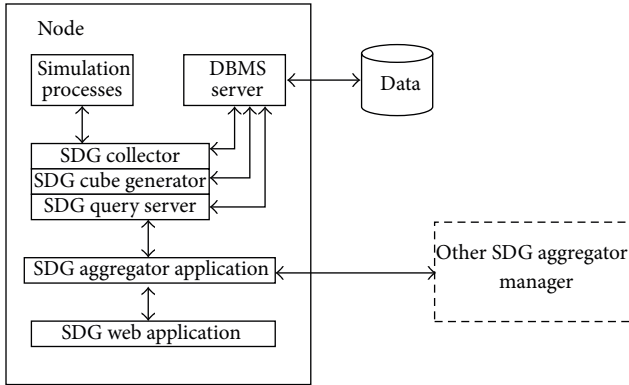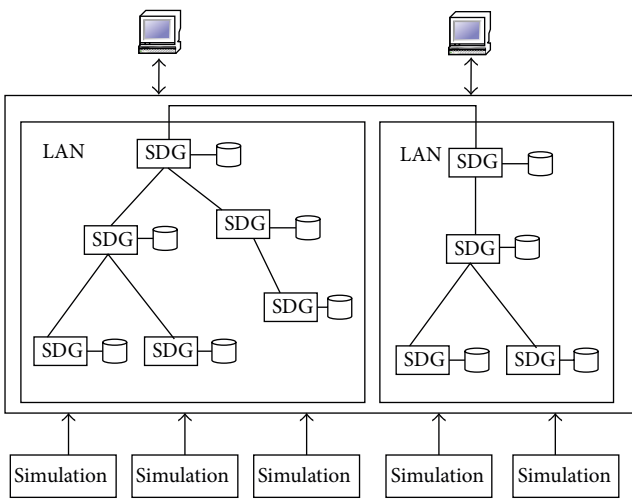
FIGURE 8: Multitier architecture of local node.

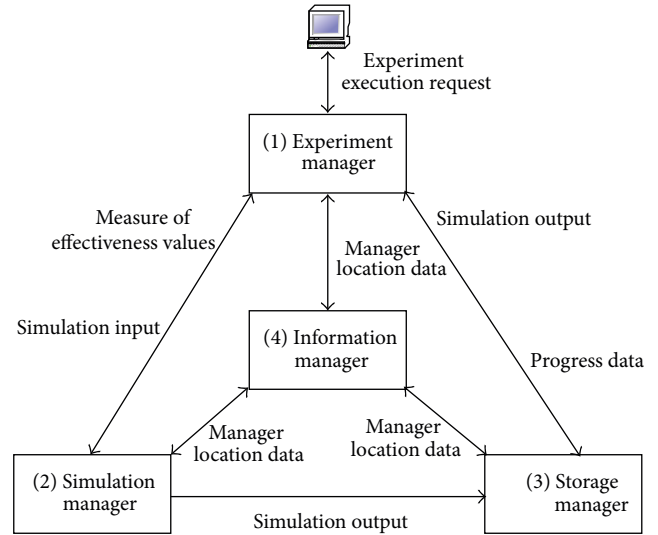FIGURE 9: Hierarchical architecture of SDG.

FIGURE 10: High-level architecture of Scalarm.

data source manager stores actual data, and the top-level manager provides a unified entrance for accessing the simulation data. The user firstly submits a data query task to the top-level manager; then the task is decomposed into subtasks, which will be assigned to other managers. The execution results of the subtasks are aggregated from bottom to top, and the final result is delivered by the top-level manager.

SDG accommodates big simulation data with scalable storage and analysis. It takes full advantage of grid computing technology and retains some fine features from the database world, such as descriptive language interface. Compared with the most popular big data framework, Hadoop, SDG provides a relatively simple concept. Furthermore, SDG is very suitable for transregional data management, which is common in joint military experiments.

4.2. Scalarm. Scalarm is a complete solution for conducting data farming experiments [27, 30, 91]. It addresses the scalable problem that is outstanding in a large-scale experiment which uses HPC to execute constructive and faster-than-real-time simulations. Scalarm manages the execution phase of data farming, including experiment design, multiruns of the simulation, and results analysis.

The basic architecture is designed as "client-master-worker" style. The master components organize resources and receive jobs from the client, and the worker components execute the actual simulation. Figure 10 presents the high-level overview of the architecture.

Experiment Manager (master component) is the core of Scalarm. It handles the experiment execution request from the user and the scheduling of simulation instances using Simulation Manager (worker component). It also provides the user with interfaces for viewing progress and analyzing results. Simulation Manager wraps the actual simulation applications and can be deployed in various computing resources, for example, Cluster, Grid, and Cloud. Storage Manager (another kind of worker component) is responsible

distributed simulation data (i.e., staying on (or near) the simulation nodes). The data analysis is also distributed so that HPC resources are reused after the simulation is complete. Therefore, it is not necessary to move the logged data across the network, because SDG sends only small results to the central site. Analysis results are aggregated via a hierarchical structure managed by SDG. This design guarantees scalable simulation and data management. A new computer node can be simply added into SDG to satisfy the growth in data size.

Figure 8 illustrates the detailed architecture of a local SDG node. The original data from a simulation are collected and saved in a rational database through the SDG Collector module. In a joint experiment based on JSAF, a plug-in within the simulation federate intercepts RTI calls and sends RTI data to the SDG Collector using network sockets. Then the SDG Cube Generator extracts facts to populate more tables, which represent user views as multidimension cubes. The SDG Query Server returns cube references for query. The SDG Aggregator receives cube objects returned from other nodes and combines them into a new one. The data query is initiated by the Web Application.

As a distributed system, SDG consists of three kinds of managers: top-level, worker, and data source (Figure 9). The

TABLE 3: Comparison of two MS big data platforms.

| Platform	Application characteristics	Areas of focus	Technical architecture	Advantages	Disadvantages
SDG	Real-time; distributed; virtual and constructive	Data collection, storage, and query	Grid	Scalable; supports heterogeneous simulation	Deployment is not flexible; complex data processing is not supported
Scalarm	Super real-time; constructive	Simulation job scheduling; data storage	SOA, cloud	Scalable; flexible; support heterogeneous resources	Complex simulation such as distributed simulation is not considered; high performance data analysis is not supported
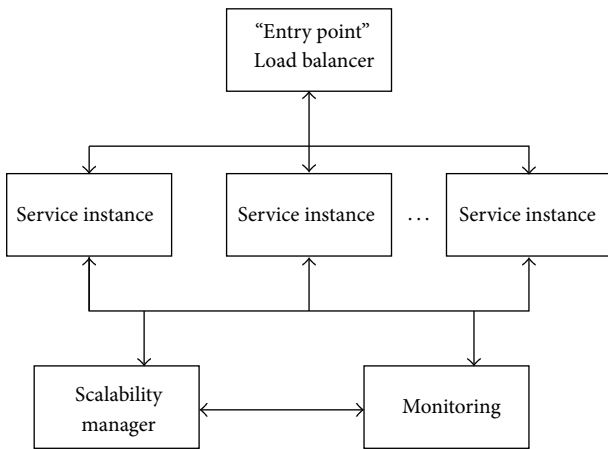
FIGURE 11: Structure of self-scalable service.

for the persistence of both simulation results and experiment definition data using a nonrelational database and file system. It is implemented as a separate service for flexibility and managing complexity. Storage Manager can manage large amounts of distributed storage with load balancing while providing a single access point. Information Manager maintains the locations of all other services, and its location is known by them. A service needs to query the location of another service before accessing. Therefore Scalarm is also a Service Oriented Architecture (SOA) system.

Two important features are supported to improve resource utilization: load balancing and scaling. Not only worker components but also master components are scalable in Scalarm. For these purposes, more components are added as master or worker to create a self-scalable service. Figure 11 illustrates the function structure. Here, a service instance can be an Experiment Manager, Simulation Manager, or Storage Manager. The load balancer forwards incoming tasks to service instances depending on their loads. The monitoring component collects workload information from each node. The Scalability Manager adjusts the number of service instances according to the workload level, so that scaling rules could be satisfied. The scaling rules are predefined by expert knowledge.

Scalarm considers both simulation and storage requirements in the data farming experiment. It supports heterogeneous resources and provides massive scalability. Other

features, such as experiment management, statistical analysis, and service reliability, are also available. As an emerging open source platform, Scalarm promises to be a foundation of future large-scale data farming projects.

4.3. Summary. The two platforms presented in this section have different focuses because of their different application backgrounds and objectives. Neither one of them can cover all requirements of an experiment. In addition, each has its respective advantages and disadvantages in implementation. Table 3 shows a comparison of the two systems.

5. Challenges and Possible Solutions

Both theory and technology of MS big data have made certain advancements; but there are still some challenges that can be identified from ongoing published research. The following challenges provide directions for future work.

(1) *Bigger Simulation and Data.* More simulated entities and more complex models will be supported by computational resources with higher performance [20, 71, 77], and thus bigger datasets will be created. The requirements for usability, reliability, flexibility, efficiency, and other quality factors of the entire system will increase along with simulation scale. The opportunity is that we will be able to obtain more value through simulation, and military decision-making can be improved.

(2) *Unified Framework Serving Both Large-Scale Simulation and Big Data [91].* Many business platforms based on cloud computing have incorporated big data framework to enhance their services and expand their applications, such as Google Cloud Platform, Amazon EC2, and Microsoft Azure [92]. Usually MS big data is both created and processed by an HPC system; however, a complete platform serving both military simulation and big data is rather limited in number. We need an integrated platform to access the models, applications, resources, and data via a single entrance point. The experimental workflow from initial problem definition to final analysis should be automated for the military user to the greatest extent possible. Furthermore, multiple experiments should be scheduled and accessed simultaneously.

(3) *Generating Data Efficiently*. High-performance simulation algorithms and software are still insufficient [67]. Large-scale military simulation can be compute-intensive, network-intensive, and data-intensive at the same time. Therefore, it is one of the most complex distributed applications, and performance optimization is very difficult to achieve. For example, the load balancing technology needs to be reconsidered because of the great uncertainty intrinsic to military models.

(4) *Consolidate Data Processing and Analytical Ability with the Latest Big Data Technology*. There are only a few practices applying the mainstream big data methods and tools to MS big data. The new parallel paradigms such as MapReduce based on Key-Value Pair representation for MS data need further study. In addition, although open source software for big data is available, it is often designed for a common purpose and can be immature in some aspects [17]. Generally it needs to be modified and further integrated into a productive environment. For example, MS big data are usually spatiotemporal, so the data storage and query must be optimized based on open source software.

(5) *Big Data Application*. There are many new analysis methods and applications emerging from business big data, such as social network analysis, recommendations, and community detection [93]. By contrast, new applications for military simulation are limited. Military problems are still far from being well-studied, and big data provides a chance to reach a deeper understanding. Some new ideas emerging from the commercial sector can be borrowed by military analysts, and the military requirements should be investigated systematically, so that the user can make better use of MS big data. On the other hand, modeling and simulation itself can also benefit from big data. For example, the simulation data can be used to validate models or optimize simulation outputs.

(6) *Change of Mindset*. Military simulation data are generated from models which need to portray the pervasive uncertainty, and this work is still confronted with many difficulties. As a result, people often doubt the simulation result. But big simulation data is useful because it has potential value for revealing patterns, if not accurate results. Mining value from simulation data means a change of viewpoint about the simulation's purpose, that is, from prediction to gaining insight [28]. This may be the biggest challenge in the field today. As proof, the concept of data farming has been proposed for many years, but it is still not broadly applied. However, a change of mindset will advance military simulation theory and technology.

Related with the emerging technologies including web service, cloud service, modeling and simulation as a service, and model engineering, [94–98], a layered framework (Figure 12) is proposed to serve as the basis for future solutions. This framework addresses the system architectural aspects of challenges above. It provides a unified environment for whole lifecycle of military simulation experiment. The key data technology is management of resources and workloads.

This framework contains 5 layers, which are explained here in detail:

(i) The *portal* provides all users with a unified entry point for ease of access and use. The functions include user management, resource monitoring, and experiment launching. Different phases in an experiment can be linked by the workflow tool so that the process is automated. Tools supporting collaboration among users are also needed.

(ii) The *application layer* provides a set of tools to define and perform experiments. The functions include creating models, defining experiment, running simulation, and analyzing results. The important components are the simulation and data processing engines, which accept experiment tasks and access computation and storage resources.

(iii) The *service layer* includes common services used by the above applications. The computing service provides computational resources for running simulations or performing analysis. The data service responds to the requests for data storage and access. The monitor service collects workload information from the nodes for load balancing scheduling. It also collects health information used for achieving reliability. The communication service provides applications with simple communication interfaces. The service manager is a directory containing the Metainformation of the resources and service instances.

(iv) The *platform layer* provides all kinds of resources including computing, storage, and communication middleware. A resource itself could be managed by a third-party platform, such as MapReduce system or HDFS. The third-party platform can directly ensure scalability or reliability with its own merit. In any case, all system resources are wrapped by upper-level services.

(v) The *repository* includes a set of data resources, such as experiment inputs and outputs, and components used for executing simulation and analysis. The algorithms library includes parallel computation methods to implement the high performance simulation at the lowest level as illustrated in Figure 4 together with parallel data processing algorithms.

The service layer in Figure 12 is the core part of the framework. The computing service encapsulates different computing resources with unified resource objects and registers itself in the service manager. Afterwards, the simulation and data processing module in the application layer request computing resources with a unified workload model from the service manager. The service manager allocates appropriate computing services according to demand and system workload conditions. This allocation is coarse-grained of processes. Application engines can make fine-grained scheduling
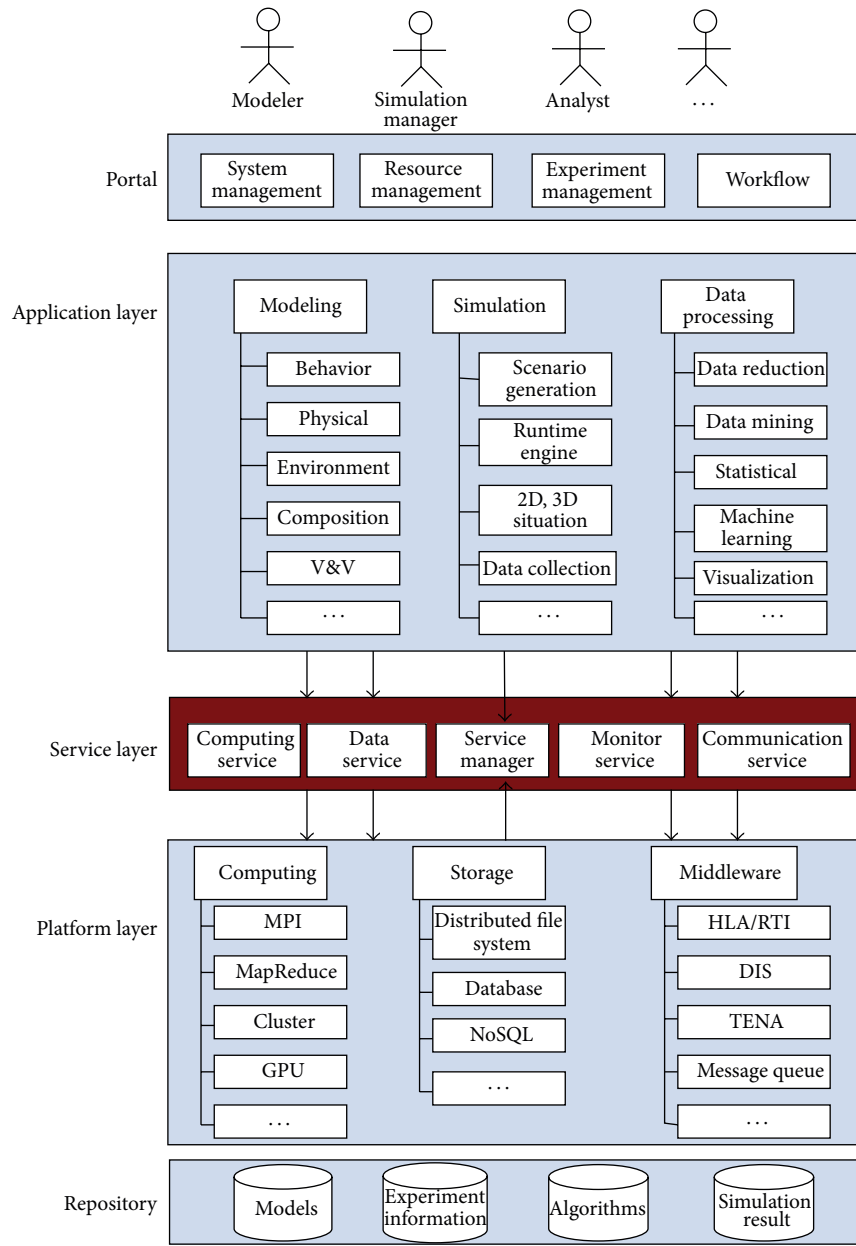
FIGURE 12: Recommended architecture of unified experiment platform.

decisions of threads. Therefore, different experiments running simulation or data analysis can be efficiently scheduled within the same framework.

The framework is designed for multitask and multiuser considering the shared experimental resources. Each experiment mainly contains two tasks: simulation execution task and data analysis task. There are two scheduling methods available. First, if the computing resource has private storage, we should keep the data at local storage and schedule the data analysis task to the same position (see Figure 13(a)). In this method, the simulation application engine should have already ensured the load balancing; thus the data analysis is automatically load balanced. Second, if the storage resource

is separated from computing resource, we can group the computing resources with the two tasks to facilitate the resource allocation (see Figure 13(b)). In this case, two groups of resources should take short distance to reduce the overhead from data transfer.

Because resources are not managed directly by the up level applications, new resources can be simply added to satisfy the increasing scale of the simulation and data. Although our framework also employs SOA with loosely coupled modules, it has two important differences from Scalarm: (1) the computing service does not access the data service directly, and both of them are accessed by encapsulated services or applications; (2) both simulation and analysis applications
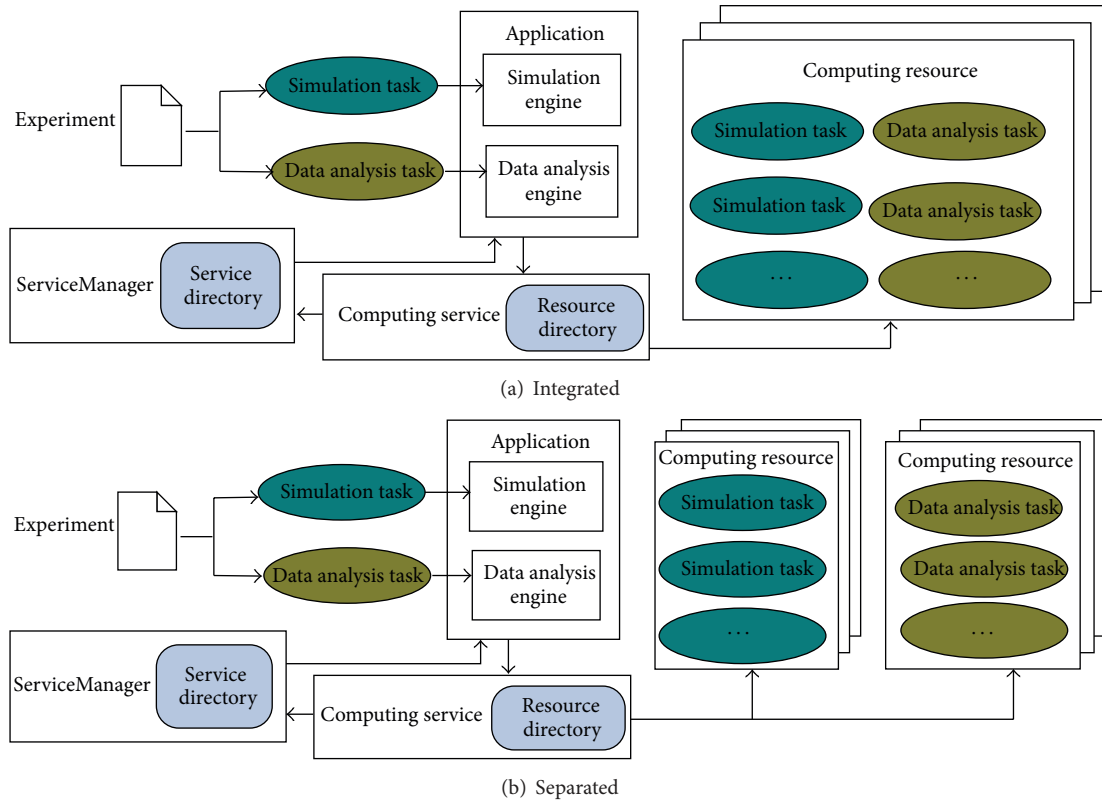
(a) Integrated

(b) Separated

FIGURE 13: Computing resource scheduling.

can request computing service and then be managed with the unified management service.

6. Conclusion

With the development of HPC technology, complex scenarios can be simulated to study military problems. This requires large-scale experiments and gives rise to the explosive growth of generated data. This paper discussed the advancement of MS big data technology, including the generation, management, and analysis of data. We also identified the key remaining challenges and proposed a framework to facilitate the management of heterogeneous resources and all experiment phases.

MS big data can change our mindsets on both simulation and big data. First, simulation was typically viewed as an approach to predict outcomes. However, prediction needs to fix many parameters for accuracy, and usually this is not feasible for military problems. However, big data can improve the analytical capability by obtaining the entire landscape of future possibilities. In fact, the idea behind big data is not novel in military simulation. Hu et al. believed that the old methodologies like data mining and data farming are consistent with big data, and the latter provides new means to resolve big simulation data requirements [67]. Second, the current big data paradigm relies on observational data to find interesting patterns [99]. The simulation experiment breaks this limitation and allows the virtual data gathered

from multiple possible futures to be our advantage. This mindset will make the best use of big data and bring more opportunities than we can imagine.

Although the big data idea in military simulation has been around for a long time, the techniques and systems are still limited in their ability to provide complete solutions. Especially, for those cases which need strict timeliness, military decision-making is posing challenges for the generation and processing of abundant data. We believe that in the near future the big data theory will further impact the military simulation community, and both analyst and decision-maker will benefit from the advancement of big data technology.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

References

[1] D. Laney, *The Importance of 'Big Data': A Definition*, Gartner, Stamford, Conn, USA, 2012.

[2] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Springer, Berlin, Germany, 2012.

[3] J. Kołodziej, H. González-Vélez, and L. Wang, "Advances in data-intensive modelling and simulation," *Future Generation Computer Systems*, vol. 37, pp. 282–283, 2014.

[4] Y. Zou, W. Xue, and S. Liu, "A case study of large-scale parallel I/O analysis and optimization for numerical weather prediction system," *Future Generation Computer Systems*, vol. 37, pp. 378–389, 2014.

[5] I. J. Donaldson, S. Hom, and T. Housel, *Visualization of big data through ship maintenance metrics analysis for fleet maintenance and revitalization [Ph.D. thesis]*, Naval Postgraduate School, Monterey, Calif, USA, 2014.

[6] O. Savas, Y. Sagduyu, J. Deng, and J. Li, "Tactical big data analytics: challenges, use cases, and solutions," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 86–89, 2014.

[7] D. Smith and S. Singh, "Approaches to multisensor data fusion in target tracking: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 12, pp. 1696–1710, 2006.

[8] D. D. Hodson and R. R. Hill, "The art and science of live, virtual, and constructive simulation for test and analysis," *The Journal of Defense Modeling and Simulation: Applications, Methodology*, vol. 11, no. 2, pp. 77–89, 2014.

[9] J. O. Miller, *Class Slides, OPER 671, Combat Modeling I*, Department of Operations Research, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, USA, 2007.

[10] J. E. Coolahan, "Modeling and simulation at APL," *Johns Hopkins APL Technical Digest*, vol. 24, no. 1, pp. 63–74, 2003.

[11] A. B. Anastasiou, *Modeling Urban Warfare: Joint Semi-Automated Forces in Urban Resolve*, BiblioScholar, 2012.

[12] R. J. Graebener, G. Rafuse, R. Miller et al., "Successful joint experimentation starts at the data collection trail-part II," in *Proceedings of the Interservice/Industry Training, Simulation & Education Conference (I/ITSEC '04)*, National Training Systems Association, Orlando, Fla, USA, December 2004.

[13] T. D. Gottschalk, K.-T. Yao, G. Wagenbreth, R. F. Lucas, and D. M. Davis, "Distributed and interactive simulations operating at large scale for transcontinental experimentation," in *Proceedings of the 14th IEEE/ACM International Symposium on Distributed Simulation and Real-Time Applications (DS-RT '10)*, pp. 199–202, Fairfax, Va, USA, October 2010.

[14] G. Wagenbreth, D. M. Davis, R. F. Lucas, K.-T. Yao, and C. E. Ward, "Nondisruptive data logging: tools for USJFCOM large-scale simulations," in *Proceedings of the Simulation Interoperability Workshop*, Simulation Interoperability Standards Organization, 2010.

[15] http://en.wikipedia.org/wiki/Data_farming.

[16] S. J. E. Taylor, S. E. Chick, C. M. Macal, S. Brailsford, P. L'Ecuyer, and B. L. Nelson, "Modeling and simulation grand challenges: an OR/MS perspective," in *Proceedings of the Winter Simulation Conference (WSC '13)*, pp. 1269–1282, IEEE, Washington, DC, USA, December 2013.

[17] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: a technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.

[18] Y. Ma, H. Wu, L. Wang et al., "Remote sensing big data computing: challenges and opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47–60, 2015.

[19] A. Castiglione, M. Gribaudo, M. Iacono, and F. Palmieri, "Exploiting mean field analysis to model performances of big data architectures," *Future Generation Computer Systems*, vol. 37, pp. 203–211, 2014.

[20] G. Horne, B. Akesson, T. Meyer et al., *Data Farming in Support of NATO*, North Atlantic Treaty Organization: RTO, 2014.

[21] http://www.smartdatacollective.com/bernardmarr/185086/facebook-s-big-data-equal-parts-exciting-and-terrifying.

[22] http://www.199it.com/archives/200418.html.

[23] V. Turner, D. Reinsel, J. F. Gantz, and S. Minton, *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*, IDC Analyze the Future, 2014.

[24] https://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/.

[25] P. A. Wilcox, A. G. Burger, and P. Hoare, "Advanced distributed simulation: a review of developments and their implication for data collection and analysis," *Simulation Practice and Theory*, vol. 8, no. 3-4, pp. 201–231, 2000.

[26] N. Q. V. Hung, H. Jeung, and K. Aberer, "An evaluation of model-based approaches to sensor data compression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 11, pp. 2434–2447, 2013.

[27] D. Król, M. Wrzeszcz, B. Kryza, Ł. Dutka, and J. Kitowski, "Massively scalable platform for data farming supporting heterogeneous infrastructure," in *Proceedings of the 4th International Conference on Cloud Computing, Girds, and Virtualization (CLOUD COMPUTING '13)*, pp. 144–149, Valencia, Spain, May-June 2013.

[28] G. Horne, S. Sanchez, S. Seichter et al., *Data Farming in Support of Military Decision Makers*, Naval Postgraduate School, Monterey, Calif, USA, 2009.

[29] G. Horne and S. Seichter, "Data farming in support of NATO operations: methodology and proof-of-concept," in *Proceedings of the Winter Simulation Conference*, pp. 2355–2363, IEEE, December 2014.

[30] http://www.scalarm.com/index.php?action=datafarming.

[31] P. Barry, J. Zhang, and M. McDonald, "Architecting a knowledge discovery engine for military commanders utilizing massive runs of simulations," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, pp. 699–704, ACM, Washington, DC, USA, August 2003.

[32] J. L. Ross, "A comparative study of simulation software for modeling stability operations," in *Proceedings of the Symposium on Military Modeling and Simulation (MMS '12)*, Society for Computer Simulation, Orlando, Fla, USA, March 2012.

[33] D. Kallfass and T. Schlaak, "NATO MSG-088 case study results to demonstrate the benefit of using Data Farming for military decision support," in *Proceedings of the Winter Simulation Conference (WSC '12)*, Berlin, Germany, December 2012.

[34] R. McCourt and K. Ng, "Enhanced network modeling in ABSNEC," in *Proceedings of the International Simulation Multi-Conference*, pp. 537–544, July 2010.

[35] J. F. O'May, E. G. Heilman, and B. A. Bodt, "Battle command metric exploration in a simulated combat environment," Final Report, Army Research Lab Aberdeen Proving Ground MD Computational And Information Sciences DIR, 2005.

[36] U. Z. Yildirim, İ. Sabuncuoglu, B. Tansel, and A. Balcioglu, "A design of experiments approach to military deployment planning problem," in *Proceedings of the Winter Simulation Conference (WSC '08)*, pp. 1234–1241, Austin, Tex, USA, December 2008.

[37] J. Logsdon, D. Nash, and M. Barnes, "One semi-automated forces (OneSAF): capabilities, architecture, and processes," in *Proceedings of the DoD M&S (Modeling and Simulation) Conference*, Orlando, Fla, USA, March 2008.

[38] R. Smith, "OneSAF: Next Generation Wargame Model," 2008, http://www.modelbenders.com/papers/RSmith_OneSAF_KIDA.pdf.

[39] C. Bouwens, O. Hasan, R. Smith, and et al, "OneSAF implementation on high performance computing systems," in *Proceedings of the Spring Simulation Interoperability Workshop*, pp. 67–77, April 2011.

[40] A. Barnett, "Using OneSAF as a Tool to Support Analysis," 2009, http://www.slideserve.com/zenevieva/using-one-semi-automated-forces-onesaf-as-a-tool-to-support-analysis.

[41] C. N. Seymour, *Capturing the full potential of the Synthetic Theater Operations Research Model (STORM) [Ph.D. thesis]*, Naval Postgraduate School, Monterey, Calif, USA, 2014.

[42] M. L. McDonald, S. C. Upton, C. N. Seymour et al., *Enhancing the analytic utility of the Synthetic Theater Operations Research Model (STORM) [Ph.D. thesis]*, Naval Postgraduate School, Monterey, Calif, USA, 2014.

[43] J. Bos, J. Graham, and J. Mitchell, "Use of M&S to support acquisition of major naval systems," in *Proceedings of the Annual Simulation Technology and Training Conference (SimTecT '10)*, pp. 73–78, Brisbane, Australia, May-June 2010.

[44] S. Boinepalli and G. Brown, "Simulation studies of naval warships using the ship air defence model (SADM)," in *Proceedings of the Annual Simulation Technology and Training Conference (SimTecT '10)*, pp. 109–114, Brisbane, Australia, May-June 2010.

[45] A. Butkus, K. Roe, B. L. Mitchell, and T. Payne, "Space surveillance network and analysis model (SSNAM) performance improvements," in *Proceedings of the IEEE DoD High Performance Computing Modernization Program Users Group Conference (HPCMP-UGC '07)*, pp. 469–473, Pittsburgh, Pa, USA, June 2007.

[46] R. Brassell, J. Nolan, and C. Stevens, "Implementation of an accelerated assessment process for the terminal high altitude area defense system: initial operational test and evaluation supporting a production decision," *International Test and Evaluation Association Journal*, vol. 33, no. 2, pp. 135–144, 2012.

[47] S. Han, K.-Y. Lam, J. Wang, K. Ramamritham, and A. K. Mok, "On co-scheduling of update and control transactions in real-time sensing and control systems: algorithms, analysis, and performance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2325–2343, 2013.

[48] B. P. Zeigler and H. S. Sarjoughian, *Guide to Modeling and Simulation of Systems of Systems*, Springer, London, UK, 2013.

[49] M. Hoogendoorn, R. Van Lambalgen, and J. Treur, "An integrated agent model addressing situation awareness and functional state in decision making," in *Agents in Principle, Agents in Practice*, vol. 7047 of *Lecture Notes in Computer Science*, pp. 385–397, Springer, Berlin, Germany, 2011.

[50] Q. Sui, Y. C. Yeo, K. Y. How et al., "An advanced rule engine for computer generated forces," in *Proceedings of the 17th Conference on Behavior Representation in Modeling and Simulation*, Providence, RI, USA, 2008.

[51] L. Feng, J.-W. Hu, Q.-J. Yin, and Y.-B. Zha, "IBMSE: an integrated behavior modeling and simulation environment," in *Proceedings of the 2nd International Conference on Computer Modeling and Simulation (ICCMS '10)*, pp. 401–405, IEEE, Sanya, China, January 2010.

[52] J. J. M. Roessingh, R. J. Merk, R. Meiland et al., "Modelling CGFs for tactical air-to-air combat training motivation-based behaviour and machine learning in a common architecture," NATO Report, 2011.

[53] K. Van den Bosch, P. Kerbusch, and J. Schram, "Modeling cultural behavior for military virtual training," in *Proceedings of the Interservice/Industry Training, Simulation & Education Conference (I/ITSEC '12)*, Orlando, Fla, USA, November 2012.

[54] M. Hoogendoorn, R. J. Merk, and J. Treur, "A decision making model based on Damasio's somatic marker hypothesis," in *Proceedings of the 9th International Conference on Cognitive Modeling (ICCM '09)*, pp. 1001–1009, Manchester, UK, July 2009.

[55] R. J. Merk, "A computational model on surprise and its effects on agent behaviour in simulated environments," in *Advances in Practical Applications of Agents and Multiagent Systems*, pp. 47–57, Springer, Berlin, Germany, 2010.

[56] S. Pickl, S. Meyer-Nieberg, and J. Wellbrink, "Reducing complexity with evolutionary data farming," *SCS M&S Magazine*, no. 2, pp. 47–53, 2012.

[57] S. D. Chi, J. Y. You, H. C. Jung, J.-S. Lee, and J.-I. Kim, "Fames: fully agent-based modeling & emergent simulation," in *Proceedings of the Spring Simulation Multiconference (SpringSim '09)*, p. 29, San Diego, Calif, USA, March 2009.

[58] S. Sun, X. Lin, J. Pan, and Y. Zhou, "Decision-making and planning of computer generated forces," in *Proceedings of the IEEE International Conference on Computer Science and Automation Engineering (CSAE '12)*, vol. 1, pp. 476–480, IEEE, Zhangjiajie, China, May 2012.

[59] S. K. Lee, C. G. Yoo, J. C. Park et al., "Autonomous intelligent simulation model based on CGF for military training and analysis," in *Proceedings of the 10th WSEAS International Conference on Information Security and Privacy (ISP '11)*, pp. 118–123, World Scientific and Engineering Academy and Society (WSEAS), Jakarta, Indonesia, December 2011.

[60] X. Tan, S. Lai, W. Wang, and M. Zhang, "Framework of wargame CGF system based on multi-agent," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC '12)*, pp. 3141–3146, IEEE, Seoul, Republic of Korea, October 2012.

[61] Y.-P. Yao and G. Liu, "High-performance simulation computer for large-scale system-of-systems simulation," *Journal of System Simulation*, vol. 23, no. 8, pp. 1617–1623, 2011 (Chinese).

[62] J. Santiago, R. Smith, J. Li, and D. Priest, "Interactive HPC-enabled LVC S&T experimentation environments," in *Proceedings of the DoD High Performance Computing Modernization Program Users Group Conference (HPCMP UGC '10)*, pp. 333–339, IEEE, Schaumburg, Ill, USA, June 2010.

[63] C. Bouwens, O. Hasan, S. Lopez et al., "OneSAF implementation on high performance computing systems," in *Proceedings of the Spring Simulation Interoperability Workshop (SIW '11)*, Simulation Interoperability Standards Organization, Orlando, Fla, USA, April 2011.

[64] C. Bouwens, A. Henninger, and A. Paschal, *OneSAF As a Simulation Service using High Performance Computing*, Contributions to DoD Mission Success, High Performance Computing Modernization Program, 2012.

[65] T. D. Gottschalk, K.-T. Yao, G. Wagenbreth, R. F. Lucas, and D. M. Davis, "Distributed and interactive simulations operating at large scale for transcontinental experimentation," in *Proceedings of the IEEE/ACM 14th International Symposium on Distributed*

Simulation and Real Time Applications (DS-RT '10), pp. 199–202, IEEE, Fairfax, Va,USA, October 2010.

[66] R. T. Kouzes, G. A. Anderson, S. T. Elbert, I. Gorton, and D. K. Gracio, "The changing paradigm of data-intensive computing," *Computer*, vol. 42, no. 1, pp. 26–34, 2009.

[67] X. F. Hu, X. Y. He, and X. L. Xu, "Simulation in the big data era—review of new ideas and new theories in the 81st Academic Salon of China Association for Science and Technology," *Scientia Sinica Informationis*, vol. 44, no. 5, pp. 676–692, 2014 (Chinese).

[68] R. E. De Grande and A. Boukerche, "Dynamic balancing of communication and computation load for HLA-based simulations on large-scale distributed systems," *Journal of Parallel and Distributed Computing*, vol. 71, no. 1, pp. 40–52, 2011.

[69] B. Möller, F. Antelius, T. W. van den Berg et al., "Scalable and embeddable data logging for live, virtual and constructive simulation: HLA, link 16, DIS and more," in *Proceedings of the Fall Simulation Interoperability Workshop*, 11F-SIW-055, Simulation Interoperability Standards Organization, 2011.

[70] B. Möller, F. Antelius, R. Brunton, T. W. van den Berg, and R. R. Witberg, "Toward a data logging data interchange format: use cases and requirements," in *Proceedings of the Fall Simulation Interoperability Workshops (SIW '11)*, Orlando, Fla, USA, September 2011.

[71] K. T. Yao, D. M. Davis, R. F. Lucas, and P. Amburn, "Supercomputing's role in data problems and its contribution to solutions," *The ITEA Journal of Test and Evaluation*, vol. 27, no. 3, 2006.

[72] W. Li, P. Ma, and M. Yang, "Research on HLA-based distributed data collection techniques," *Journal of Sichuan University: Engineering Science Edition*, vol. 44, no. 3, pp. 137–140, 2012 (Chinese).

[73] D. M. Davis and R. F. Lucas, *Joint Experiment on Scalable Parallel Processors (JESPP) Parallel Data Management*, Information Sciences Institute, University of Southern California, 2006.

[74] Y. Wu and G. Gong, "A fully distributed collection technology for mass simulation data," in *Proceedings of the 5th International Conference on Computational and Information Sciences (ICCIS '13)*, pp. 1679–1683, IEEE, Shiyang, China, June 2013.

[75] M. Parashar, "Big data challenges in simulation-based science," in *Proceedings of the 6th International Workshop on Data Intensive Distributed Computing (DIDC '14)*, pp. 1–2, ACM, Vancouver, Canada, June 2014.

[76] L. Gong, B. Huang, and Z. Liu, "Research on technique of data storage in mass battle simulation," *Computer & Digital Engineering*, vol. 40, no. 2, pp. 52–55, 2012 (Chinese).

[77] K. T. Yao, R. F. Lucas, C. E. Ward et al., "Data analysis for massively distributed simulations," in *Proceedings of the Interservice/Industry Training, Simulation & Education Conference (I/ITSEC '09)*, Orlando, Fla, USA, 2009.

[78] K. T. Yao, C. E. Ward, and D. M. Davis, "Data fusion of geographically dispersed information: experience with the scalable data grid," *ITEA Journal*, vol. 32, pp. 86–94, 2011.

[79] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[80] J. Zhang, J. Bala, P. S. Barry, T. E. Meyer, and S. K. Johnson, "Mining characteristic rules for understanding simulation data," in *Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '02)*, pp. 381–386, IEEE, Washington, DC, USA, November 2002.

[81] J. F. O'May, E. G. Heilman, and B. A. Bodt, *Battle Command Metric Exploration in a Simulated Combat Environment*, Army Research Lab, Computational and Information Sciences DIR, Aberdeen, Md, USA, 2005.

[82] D. Dor and Y. Minkov, "Broadening quantitative analysis of distributed interactive simulation with data mining functionalities," in *Proceedings of the Interservice/Industry Training, Simulation & Education Conference (I/ITSEC '12)*, National Training Systems Association, Orlando, Fla, USA, November 2012.

[83] Y. Minkov, D. Deutch, M. Novoselski et al., "Process oriented development for effective data analysis of military distributed simulation," in *Proceedings of the Spring Simulation Interoperability Workshop (SIW '11)*, pp. 48–56, SISO, Orlando, Fla, USA, April 2011.

[84] Y. Yin, G. Gong, and L. Han, "Experimental study on fighters behaviors mining," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5737–5747, 2011.

[85] Y. Yin, "Data mining and control approach to formation flying," in *Computer Science for Environmental Engineering and EcoInformatics*, pp. 321–326, Springer, Berlin, Germany, 2011.

[86] L. D. Acay, "Semi-automated causal analysis in simulation based military training," in *Proceedings of the Simulation Conference and Exhibition (SimTecT '10)*, pp. 353–358, Brisbane Convention Centre, May 2010.

[87] S. M. Sanchez and T. W. Lucas, "Exploring the world of agent-based simulations: simple models, complex analyses," in *Proceedings of the 34th Conference on Winter Simulation: Exploring New Frontiers*, vol. 1, pp. 116–126, San Diego, Calif, USA, December 2002.

[88] B. Chandrasekaran, J. R. Josephson, J. O'May, E. Heilman, and R. C. Kaste, "Mining simulation data for insights about a decision space: application to an urban combat COA," in *Proceedings of the Enabling Technologies for Simulation Science VIII*, vol. 5423 of *Proceeding of SPIE*, August 2004.

[89] R. Kaste, J. O'May, E. Heilman et al., "From simulation to insights: experiments in the use of a multi-criterial viewer to develop understanding of the COA space," in *Proceedings of the CTA Conference*, 2003.

[90] C. J. Clark and P. Hallenbeck, "Data and event visualization," in *Proceedings of the ITEA Live-Virtual Constructive Conference*, El Paso, Tex, USA, 2009.

[91] D. Król, *Massively self-scalable platform for data farming [Ph.D. thesis]*, AGH University of Science and Technology, 2014.

[92] M. Alam and K. A. Shakil, "Recent developments in cloud based systems: state of art," CoRR, http://arxiv.org/abs/1501.01323.

[93] J. Manyika, M. Chui, B. Brown et al., *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, 2011.

[94] B. P. Zeigler and L. Zhang, "Service-oriented model engineering and simulation for system of systems engineering," in *Concepts and Methodologies for Modeling and Simulation: A Tribute to Tuncer Ören*, L. Yilmaz, Ed., Simulation Foundations, Methods and Applications, pp. 19–44, Springer, Berlin, Germany, 2015.

[95] E. Cayirci, "Configuration schemes for modeling and simulation as a service federation," *Simulation*, vol. 89, no. 11, pp. 1388–1399, 2013.

[96] X. Song, Y. Ma, and D. Teng, "A load balancing scheme using federate migration based on virtual machines for cloud simulations," *Mathematical Problems in Engineering*, vol. 2015, Article ID 506432, 11 pages, 2015.

[97] E. Cayirci, "Modeling and simulation as a cloud service: a survey," in *Proceedings of the Winter Simulation Conference*, pp. 389–400, IEEE, Washington, DC, USA, December 2013.

[98] Y. Wu, X. Song, and G. Gong, "Real-time load balancing scheduling algorithm for periodic simulation models," *Simulation Modelling Practice and Theory*, vol. 52, no. 1, pp. 123–134, 2015.

[99] S. M. Sanchez, "Simulation experiments: better data, not just big data," in *Proceedings of the Winter Simulation Conference*, pp. 805–816, IEEE Press, Savanah, Ga, USA, December 2014.

Advances in
Operations Research

Advances in
Decision Sciences

Journal of
Applied Mathematics

Algebra

Journal of
Probability and Statistics

The Scientific
World Journal

International Journal of
Differential Equations

International Journal of
Combinatorics

Advances in
Mathematical Physics

Journal of
Complex Analysis

Journal of
Mathematics

Mathematical Problems
in Engineering

Abstract and
Applied Analysis

Discrete Dynamics in
Nature and Society

International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces

International Journal of
Stochastic Analysis

Journal of
Optimization