Tactical Big Data Analytics: Challenges, Use Cases, and Solutions

Onur Savas, Yalin Sagduyu, Julia Deng, and Jason Li
Intelligent Automation, Inc.
Rockville, MD 20855, USA
{osavas,ysagduyu,hdeng,jli}@i-a-i.com

ABSTRACT

We discuss tactical challenges of the Big Data analytics regarding the underlying data, application space, and computing environment, and present a comprehensive solution framework motivated by the relevant tactical use cases. First, we summarize the unique characteristics of the Big Data problem in the Department of Defense (DoD) context and underline the main differences from the commercial Big Data problems. Then, we introduce two use cases, (i) Big Data analytics with multi-intelligence (multi-INT) sensor data and (ii) man-machine crowdsourcing using MapReduce framework. For these two use cases, we introduce Big Data analytics and cloud computing solutions in a coherent framework that supports tactical data, application, and computing needs.

Keywords

Big Data, Cloud Computing, Analytics, Algorithms, Tactical Environment

1. INTRODUCTION

While it has become apparent that data is being collected at unprecedented rates thanks to a wide range of high-resolution high-throughput sensors, it has also become apparent that suitable algorithms and tools to satisfactorily analyze Big Data are largely missing. In the tactical domain, this challenge is even more amplified, where scientifically collected tactical data has missing links, is mostly unstructured and heterogeneous, and involves different levels of completeness and standardization.

Current DoD systems and processes for managing and analyzing the tactical information cannot be effectively scaled to meet the challenge of ever growing data. At the same time, the tools, algorithms, and data management techniques that we can borrow from the commercial world do not directly apply to the needs of the applications in the DoD domain. For example, the envisioned *Naval Tactical Cloud* consists of a group of clouds that reside at geographically large distances such as in the Pacific Shore, the Atlantic Shore, and Carrier Groups in both the Atlantic and Pacific Oceans [1]. Therefore, fundamentally new approaches of analysis and data management pertaining to mission decision cycles are needed.

Cloud computing is envisioned to be beneficial for Big Data analytics. However, consolidating all computational and storage resources in Big Data centers (as is mostly done in the commercial domain) is not efficient (and often not feasible) in tactical domains. The three most important challenges to consolidated data centers and tactical Big Data analytics are given as follows.

- **The tactical cloud will most likely be bandwidth limited and possibly out of network for long periods of time.** Bandwidth is very expensive in challenging environments, for example, in open sea or in countries where infrastructure is limited. Routine operations such as data replications may not be possible.
- **Security.** The net-centric cloud allows data to be stored in a distributed manner, and some of the data services are provided by mobile computing platforms (e.g., Humvees). Moreover, some data centers/servers are located in a place close to the sensors in an unfriendly environment. The data server/center might not be under the full control of a trusted authority; instead, it may be rented, provided, or maintained by an untrusted unit.
- **In tactical environments, the roles can change very quickly.** In the commercial domain, it is clear who the producers and the consumers are. In a Battlespace environment, those roles can change very quickly. A Warfighter can be the consumer of data and analytics for long periods of time, but depending on the mission, the Warfighter might take the role of data producer via portable sensors.

2. TACTICAL BIG DATA CHALLENGES

We distinguish the main characteristics of the tactical Big Data problem as follows.

2.1 Applications and Data

- **Intelligence, Surveillance and Reconnaissance (ISR) applications.** Some of the ISR applications are common with commercial ones such as Intrusion Detection and Anomaly Detection. However, some of the ISR applications are unique to tactical environments such as Target Tracking and Localization (TTL) and Persistent Surveillance. They often come with stringent delay requirements and involve real-time or near real-time objectives.

Big Data Analytics Workshop in conjunction with ACM Sigmetrics 2013
June 21, 2013.

- **Mission-driven goals.** Goals in the tactical Big Data problem are driven by strict mission needs, rather than by economics as in commercial applications. This introduces additional constraints into the problem space, such as hierarchical order of data sources, processors, and users (beyond the server-client paradigm in commercial applications).

- **Heterogenous data sources.** Most of the DoD data is unstructured, such as signals, text, image and video with different standardization. The data is obtained by a variety of sensors (e.g., LIDAR, RADAR, hyper spectral imaging (HSI), electro-optical(EO), infrared (IR), videos) over large geographical areas with different resolutions, completeness, and uncertainty. Most of the time the data is not transactional (e.g., sensor data is not like purchase data and it contains high degrees of uncertainty).

- **Uncertain/incomplete/noisy data.** The uncertainties can arise from various inaccuracies and they should be represented in the data structure. Fuzzy approaches are not sufficient to address the uncertainty in the data collection (e.g., low SNR). Methods that address uncertainties arising from algorithms (e.g., sub-optimal learning algorithms), logical inconsistencies in the model (e.g., conflicting schemas) and scalable novel methods are required.

- **Stringent security requirements.** Having military data co-located in the same virtual environment as other commercial offerings may not satisfy DoD's stringent security requirements (such as protection against data theft and corruption attacks).

2.2 Computing Architecture

- **Limited networking bandwidth.** The bandwidth is extremely expensive, for example, in challenging (highly contested) areas such as open sea and enemy zones. In addition, military radios (e.g., in airborne networking) are inherently subject to a harsh communication environment (e.g., wireless fading, multipath, mobility) and jamming/eavesdropping attacks.

- **Heterogenous processing capabilities.** Tactical networks consist of diverse users, ranging from smart phones to data centers, co-existing and interacting in the same environment of data collection, processing, and delivery.

- **Different roles.** In a tactical environment, a user may assume different roles at different times: it may be a producer (providing data), a processor (providing computing capability), or a consumer (demanding tasks). This goes beyond the server-client paradigm in commercial applications.

- **Distributed system requirements.** Sensing, storage and computation units in tactical applications are typically not co-located but rather distributed over a geographical area.

End-users in a tactical environment includes actors such as a Warfighter and an intelligent analyst, each with distinct properties. The Warfighter is characterized by uncertain query vs. uncertain data, and only knows that he/she is near a landmark, but is not sure about the exact location; and the locations of the landmarks in the database are in low spatial accuracy and some of them are no longer there. The Warfighter properties can be summarized as:

- low computing capability (probably a PDA/tablet).
- real-time requirements.
- limited technical expertise.
- simple tasks (queries) (e.g., asks questions such as "Tell me about the [bridge/valley] in front of us." or "Has there been an [IED attack/explosion] [around this area] [in the last 3 months]?"

On the other hand, the intelligence analyst exceeds the computing capabilities of the Warfighter. The intelligence analyst properties can be summarized as:

- high computing capability (with powerful data centers / clouds).
- batch processing.
- experienced/trained.
- complex interactions with the system (e.g., can use specialized query languages).

3. TACTICAL BIG DATA USE CASES

3.1 Use Case 1: Big Data Analytics using Multi-INT Sensor Data

As a first use case, we consider the Big Data analytics using multi-intelligence (multi-INT) sensor data. The main challenge is that the analysis should be carried out via multi-level data fusion using geographically dispersed data. The data sources are from full motion video (FMV), imagery, wide area surveillance, EO/IR, RADAR, and human intelligence (HUMINT). In addition, cyber domain sensors in terms of crawlers can also be used to collect text data from social media, news, blogs, and comments. Since the sensors are in geographically distributed locations, the sensor data can be stored in (i) a large data archive for retrieval and extraction (e.g., in a cloud), (ii) kept at an aggregation node (e.g., a mobile gateway), or (iii) remain close to the sensors and triggers provided for data distribution (e.g., sensors themselves).

Intelligent Automation, Inc. (IAI), in various DoD sponsored projects[1], addresses some of the challenges pertaining to joint analytics, distributed storage, and content distribution in challenging (e.g., bandwidth limited) environments. We are envisioning a system, where end-users such as Warfighters with their PDAs/tablets interact with the system using natural language questions. Consider an example, where a Navy ship has pulled into a foreign port in a country that has strained relationships with the US. The goal is to predict whether a social uprising will happen or not (in the next few days or so). A possible question is *"Will there be [a social uprising] around [the port] in the [next two days]?"* As much as the question is simple and

[1]An up-to-date project list is available at http://i-a-i.com.

crisp, the analysis requires querying and running analytics across geographically distributed data stores. In addition to the volume of data, the variety of the data makes the analytics more complex. To address these challenges, we are building a system that consists of five layers, namely Applications Layer, Semantic Layer, Analytics Layer, Storage Layer, and Distribution Layer. Based on the Warfighter's question, we walk through the layers, as follows.

Application Layer: The interactions with the end users are enabled via Ozone Widgets [6]. The Ozone Widgets are customizable open-source light weights web applications that assemble the tools needed to accomplish a broad class of tasks and enable those tools to communicate with each other. The visualization is also an important component in this layer.

Semantic Layer: This layer digests natural language questions, determines the category of the questions, and parses and analyzes the questions by using UC Berkeley FrameNet [7]. Then the query abstract is prepared. The query abstract process involves populating the query when possible. In the case of social uprising, a relevant social movement ontology could be used. The main challenge in this layer is to achieve semantic interoperability across diverse datasets.

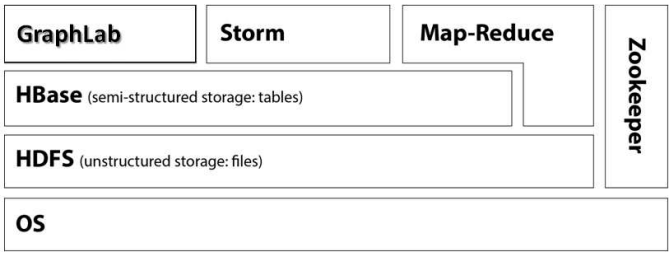

Figure 1: IAI Big Data analytics architecture.

Analytics Layer: This layer is responsible for Big Data analytics, computation, and the bulk of the processing. In general, both batch processing and near-real-time analytics are expected to be supported. The batch processing can be wrapped inside Hadoop Core, is executed as MapReduce jobs and is supported by other Hadoop ecosystem components such as Pig, HBase, and Hive. The graph analytics can be supported by GraphLab [8], which uses a shared memory architecture. Storm [9] can be leveraged to support short Command and Control decision cycles. As an example, consider full motion video depicting some activity and Twitter messages, which contain negative sentiment against US deployment in the particular port. From the video, we can extract various features depicting suspicious activities. Then we use topic modeling algorithms such as Latent Dirichlet Algorithm [10] to classify the activity and predict whether there is a threat or not. Similarly, one could check whether the Twitter messages are becoming viral or not using, for example, graph analytics. Of course, the challenge is to build the capability of running these algorithms using vast amount of data limiting false alarm rates. The high-level architecture we are developing at IAI to support the analytics capabilities is shown in Figure 1.

Storage Layer: As mentioned before the data does not need to be stored in a data center and can sometimes reside close to the sensor and at aggregate nodes. Then the challenge is to support content replication and maintain the desired consistency and availability across data stores. Also distributed indexing such as distributed hash tables should support data discovery in a reasonable amount of time.

Distribution Layer: This layer is mainly responsible for content distribution. After documents are submitted by content providers, the documents are first stored in an origin server. The content is then replicated on other surrogates (caching servers) in several situations, such as distribution tasks, content provider's preferences, content access statistics, or load balancing requirements. In all these situations, content and relevant documents are copied to one or more surrogates in other regional networks to speed up data access for end users.

3.2 Use Case 2: Man-Machine Crowdsourcing using MapReduce Framework

Crowdsourcing is becoming an effective mechanism to accomplish tasks online. The evolving tactical cloud architecture is well suited to allow crowdsourcing to be used during crisis response periods. However, the implementation of crowdsourcing as a distributed analytic capability for a commander during a crisis response mission is expected to be different than the commercial use of this technique. In particular, to accommodate crowdsourcing (e.g., for disaster and crises responses) it is necessary to receive from both humans and machines with different levels of capabilities and should be carefully combined for reliable task execution.

To enable and support effective use of crowdsourcing during crisis response periods, Intelligent Automation, Inc. envisions a "Crowdsourcing for Crisis and Disaster Applications (CrowdApp)" system. The architecture is inspired by CrowdForge [2]. CrowdApp, in broad terms, is a framework for accomplishing complex tasks from both human intelligence and machine analytical functions. The CrowdApp framework populates workflows from small tasks that can be combined and nested, to address disaster and crisis applications including data collection and recognizing social changes or activity.

Both human intelligence and machine analytic problem solving processes follow the map reduce construct, where intermediate $< key, value >$ pairs are fed to either "reduce"-tasked humans (workers) or compute nodes. For humans, the map and reduce steps are populated to solve cognitively loaded and pattern recognition focused tasks, while the tasks for machines focus on computationally loaded tasks. The individual results are automatically reduced by appropriately combining them. CrowdApp automatically manages available resources by optimally scheduling task primitives based on (i) human expertise and cognitive load, and (ii) machine availability, capability, and reliability, hence taking full advantage of data locality.

Using Figure 2, some of the advantages of our framework can be enumerated as:

- **CrowdApp allows complex tasks to be submitted (#1) and automatically populates a workflow with optimal scheduling of primitive tasks (#2):** Previous works mainly focus on simple tasks such as image labeling or judging the relevance of search results [3]. Here we envision a more general purpose framework for accomplishing complex tasks such as social uprising detection(e.g., situational awareness regarding a foreign embassy) or disaster relief (e.g., flood

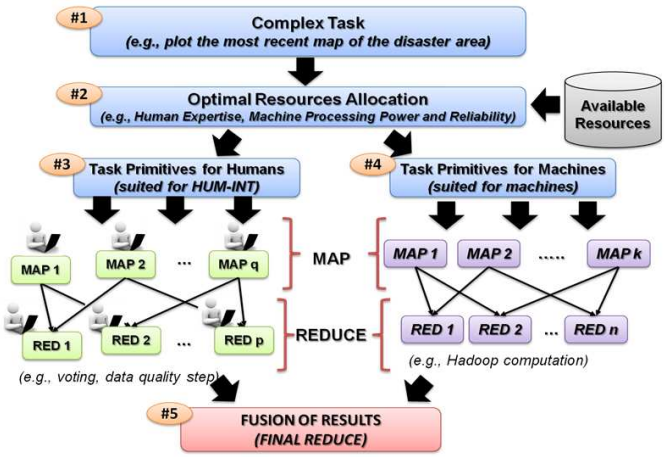

Figure 2: The envisioned CrowdApp architecture.

area mapping). Our framework allows dynamic partitioning so that workers (human resources in the system) themselves can decide a task partition, with their results in turn generating new subtasks (as opposed to the task designer asking for fully specified partitions beforehand). CrowdApp also allows multi-level partitions in which a task can be broken up by more than one partition.

We can run automatic node discovery that leverages advanced social media analytics based on graph-based community detection [4]. This can be followed by optimal scheduling for both humans and machines based on the knowledge of available resources such as (i) human expertise, cognitive load, and (ii) machine availability, processing power, and reliability.

- **CrowdApp admits crowdsourcing for Human-Intelligence Tasks (HIT) with automated reduce process (#3):** We follow a three step process (partition, map, and reduce) to crowdsource complex tasks to accomplish high quality results. In particular, using the partition step, a larger task is broken down into discrete subtasks. In map tasks, a specified subtask is processed by one or more workers, and finally, in reduce tasks, the results of multiple workers tasks are merged into a single output via voting process in general. This three-step process admits seamless management of subtasks and flows between tasks.

- **CrowdApp allows crowdsourcing for Machine Analytic Functions based on MapReduce (#4):** While the most popular implementation of the map reduce construct is Hadoop in a cloud environment, this two-step problem solving methodology can be applied to other processing nodes as well. Therefore, we consider all processing units, especially those closer to the Warfighter and disaster area to be available for distributed machine analytic computation so that the system can effectively exploit data locality. If a data center is available, i.e., a traditional cloud, then we can use Hadoop and its ecosystem to solve the primitive tasks assigned to this data center. If less powerful nodes are available, one option is to use Sector/Sphere [5] to carry out distributed task primitives, which allows parallel data processing with very simple APIs. Sector/Sphere is can also operate in a wide area network (WAN) setting that is suited to the tactical cloud.

- **CrowdApp supports a data fusion workflow to use both human and machine responses (#5):** The tasks from #1-#4 output the results from (i) humans and (ii) machine analytic functions. However, the data quality, trustworthiness, confidence and information value of the outputs differ depending on whether the subtask is processed by humans or machines. In the data fusion step, we consider the human and machine factors that go into the decision process. For example, there is no concept of expertise in machines or the humans produce different results based on their cognitive workload, expertise, and even time of the day. Systematic study of these factors along with other cognitive factors aims to produce a set of data fusion rules/functions based on well-studied theories of data fusion such as Bayesian and fuzzy logic that automatically "reduces" the outputs from partial decision processes.

The overall system can be implemented by leveraging available APIs and other tools such as Ushahidi or Amazon Mechanical Turk [3] under representative crisis and disaster scenarios such as detection of social uprising, intelligence report writing, and disaster area mapping.

4. CONCLUSION/DISCUSSION

We envision, as tactical clouds mature, that more services and analytical capabilities will be enabled. One application is automated sensor planning (or sensor management) based on shared situation awareness (SA). In other words, sensors can be dynamically tasked (or re-tasked) based on the latest status of information requirements and on-line analytic predictive processing (OLAP). In particular, a systematic approach based on cloud computing could provide scalable data mining/analysis algorithms as well as tools and platforms for ingesting real-time sensor data (e.g., technical, semantic, unstructured) for shared SA and predictive processing, and drive the sensor planning loop.

5. REFERENCES

[1] Capt. Donald Harder, "Moving Navy Command and Control into the Future," January 2013.
[2] A. Kittur, et al., "CrowdForge: Crowdsourcing Complex Work," *in Proc. of USDI*, 2011.
[3] Amazon Mechanical Turk, *https://www.mturk.com.*
[4] A. Clauset, et al., "Hierarchical structure and the prediction of missing links in networks," *Nature*, 453:98, 2008.
[5] Sector/Sphere, *http://sector.sourceforge.net/.*
[6] Ozone Platform, *http://owfgoss.org/.*
[7] FrameNet Project, *https://framenet.icsi.berkeley.edu/fndrupal/.*
[8] Y. Low, et al., "GraphLab: A New Framework for Parallel Machine Learning," *in Proc. UAI*, 2010.
[9] Storm, *http://storm-project.net/.*
[10] D. Blei, et al., "Latent dirichlet allocation," *Journal of Machine Learning Research*, 3:993-1022, 2003.