eXsight: An Analytical Framework for Quantifying Financial Loss in the Aftermath of Catastrophic Events

Matthew Coelho and Andrew Rau-Chaplin
Faculty of Computer Science
Dalhousie University
Halifax, Canada
Email: {coelho, arc}@cs.dal.ca

***Abstract*—In this paper we explore the design of an analytical framework for quantifying financial loss in the aftermath of catastrophic events. The idea is to aggregate the thousands of exposure databases received by a single reinsurer into a giant loosely structured exposure portfolio and then use Big Data analysis technology, originally developed in the context of web-scale analytics, to rapidly perform natural but ad-hoc loss analysis immediately after an event. As in many situational analysis problems, the challenge here is to work with both categorical and geospatial data, deal with partial data often at varying levels of aggregation, integrate data from many sources, and provide an analysis framework in which analyses can be rapidly performed in the hours, days, and weeks immediately after an event.**

***Keywords*-Risk analytics; framework; exposure data; post-event; MongoDB; catastrophe; reinsurance;**

I. Introduction

Property catastrophe insurance and reinsurance companies are financial institutions that provide for the equitable transfer of the risk due to catastrophic events in exchange for a premium. In the hours, days, and weeks immediately after an event, these insurers face an acute situational analysis and management challenge. They want to immediately start to flow funds to their affected clients. These insurers need fast estimates of likely losses so they can reserve the necessary capital, and communicate their potentially changed financial situations to regulators, rating agencies, and stockholders.

In 2013 two-hundred and ninety-six global natural disasters (including earthquakes, floods, and hurricanes) caused a total economic loss of $192 billion dollars (USD) [1]. Of this total economic loss, $45 billion dollars was insured meaning that the financial resources required to help effect rapid recovery were available and being held in reserve by the insurance and reinsurance companies who had underwritten the risk. Immediately after an event, these insurers want to immediately start to flow funds to their affected clients however, the situation on the ground after a natural disaster is often unclear. The extent of the affected area, the intensity of the hazard, and its impact on the value and usability of buildings are all unknown. Eventually, when initial claims have been filed, the buildings repaired or replaced, and the insurance for lost use or business interruption covered, the total loss will be know. Until then, insurers and reinsurers need systems to help them estimate their losses.

Exposure data, in the context of property catastrophe insurance, refers to data that describes what is being insured and under what terms. Broadly, it consists of three types of data: 1) Location data such as latitude/longitude, street address, country, etc. 2) Physical data such as building type, construction, age, etc. and 3) Contractual data such as coverage value, limits, deductibles, and other financial terms defining the risk transfer contract.

Primary insurance companies collect exposure data from their clients (home owners and businesses), place it in exposure databases, and pass it on to reinsurance companies. Thousands of these exposure databases are collected by reinsurance companies (from their clients - the primary insurers) and are used in the pricing process. The reinsurers typically take each individual database and run it through a pricing model to produce an expected loss table (ELT) and then archive it. The individual exposure databases are currently considered to be too big and granular to be of much further use in the reinsurer's analytical pipeline.

In this paper we explore the design of an analytical framework for quantifying financial loss in the aftermath of catastrophic events. The idea is to aggregate the thousands of exposure databases received by a single reinsurer into a giant, loosely structured exposure portfolio, and then to use Big Data analysis technology, originally developed in the context of web-scale analytics, to rapidly evaluate natural, but ad-hoc, loss analysis immediately after an event. As in many situational analysis problems, the challenge here is to work with both categorical and geospatial data, deal with partial data often at varying levels of aggregation, integrate data from many sources, and provide an analysis framework that in which natural but ad-hoc analysis can be rapidly performed in the hours, days, and weeks immediately following an event.

II. Scenarios

To help illustrate the challenges of post-event analytics and to build up a set of use cases and a definition of core framework operations, this section explores scenarios

built around three recent events. While governments, non-governmental agencies, insurers, and reinsurers are all involved in post-event exposure analysis, to keep our scope manageable we will focus on analysis from a reinsurer's perspective.

A. Tōhuku Earthquake, Tsunami, and Radiation Disaster

On March 11, 2011 a magnitude 9.0 earthquake occurred just off of the coast of Japan. The earthquake also caused a subsequent tsunami whose total run-up height measured 38.9 meters, approximately the size of a 12-story building. The combination of the earthquake and tsunami severely damaged a number of the reactors in the Fukushima I Daiichi nuclear power plant, which caused a nuclear incident, leaking radiation into the surrounding environment. This earthquake was the fourth largest earthquake in recorded history, the largest in Japan[1], and resulted in 15,854 deaths and 3,203 missing persons[2] in Japan [2].

After a catastrophic event, reinsurance companies need to calculate their loss information. To do this they need to determine the boundary of the region impacted, determine a map of hazard intensities within that region, estimate mean damage ratios (MDR) maps for the impacted by building type, and estimate losses by taking into account financial terms under a variety of assumptions.

For an event like the Tōhuku event, reinsurers would overlay a variety of hazard maps, such as shakemaps for earthquake intensity, inundation maps for tsunami intensity, and wind-borne debris maps for the radiation fallout, to determine the affected area. Overlaying these maps provides a picture of the affected event area and hazard intensities sustained within it.

After constructing this affected region, reinsurers would then need to identify the impacted exposure. In areas like the US, where detailed high quality exposure data is typically available, this might be as simple as performing a geospatial query in the exposure portfolio to identify locations with lat/longs within the boundary. In a case like the Tōhuku event which involves Japanese exposure, the specific exposure would likely be unknown. In this case, aggregated exposure collected at a district or even prefecture level would need to be spatially disaggregated using data such as daytime or nighttime population numbers to produce detailed representative lat/long based exposure.

Finally, by combining the identified exposure, the MDR maps for the event, vulnerability curves by exposure type, and a financial terms simulator, an ad-hoc event specific loss model can be constructed and used. This model will provide loss estimate summaries, a breakdown of losses by a set of filterable fields, and a mapping of losses over the area. The reinsurer can then use this information to drive the early loss settlement processes, as well as to provide information to regulators, rating agencies, and other interested parties.

[1]Since instrumental recordings began in 1900

[2]As of March 8, 2012

B. 2011 Thailand Flood Disaster

While the Tōhuku Earthquake was an event that rapidly unfolded, the 2011 Thailand Floods is an example of a slow event that unfolds over months in which the issue is unknown risk hidden in the exposure data.

In 2011 heavy rains throughout Thailand, the remnants of tropical depressions Haima and Nock-Ten, and an active monsoon season caused severe flooding across the country [3]. Although Thailand has a history of flooding, this was the most expensive event resulting in approximately $45.7 billion USD in economic losses. The biggest contributor to these losses was the manufacturing sector, contributing approximately $32 billion [4]. Many companies in Thailand's manufacturing industry are hard drive manufacturers. The impact of flooding was so extreme that it interrupted the global supply chain of hard drives, driving up world prices substantially [5].

Before 2011, the average reinsurer would have told you that they had little financial exposure to Thailand floods. Thailand's manufacturers were largely insured by Japanese primary insurers who provided reinsurers with only aggregated exposure data. It was only when the early claims started trickling in that the reinsurers realized they might have a problem. But how big was the magnitude of the problem? Answering this requires an ad-hoc analysis process that combines aggregate exposure data with publicly available industry and economic data.

The first task was to estimate the commercial exposure, and generate a detailed representative exposure set. Given knowledge of the aggregate exposure and the average value of a disk manufacturing facility, an estimate on the number of facilities and their values could be obtained. Then using a description of the transportation network and estimates of daytime population (as a proxy for the spatial distribution of commercial activity) detailed representative exposure sets could be generated.

The second task was to create a crude aggregate loss model from the early claims data, industrial building vulnerability curves, and flood inundation maps. These models could be used for reserving the capital required to payout future claims. In addition, based on the detailed exposure data, spatial accumulation modeling could be performed to identify potential loss hotspots. Spatial accumulation modeling identifies the largest exposure accumulations within circles of a given radius. This helps companies model circles of maximum potential loss, which is particularly important if the event may spread.

C. Hurricane Sandy

On the evening of October 29, 2012 Hurricane Sandy, a post-tropical cyclone, made landfall near Brigantine, New

Jersey. Although a minimal hurricane as measured by the Saffir-Simpson scale, the storm covered a massive area and caused high storm surge over large parts of the coastline. The highest inundations were located in New York, New Jersey, and Connecticut, with the above ground storm surge ranging between 2-9 ft with an average of 4.5 ft in New York [6]. Although this was a low intensity event, the huge size of the affected area and the high value of the exposure in the affected region caused damages of approximately $50 billion USD, resulting in Sandy being the second most expensive hurricane in US history [7].

In the case of Hurricane Sandy, post event analysis was greatly helped by the rich and detailed nature of the available exposure data. When you know exactly what is being impacted (the exposure) you can concentrate your analysis on getting a more detailed and closer to real-time view of the evolving event. One interesting opportunity that became apparent during Hurricane Sandy was the potential for building real-time event intensity and impact maps from information gleaned from social media data. Social media is a new form of data we can look at for pre, peri, and post-event analysis. The idea is to supplement physical intensity measures with observed intensity measures using Geotagged Tweets, Instagrams, and Facebook posts as data sources. The goal is to try and build up new high quality, real-time hazard maps from on-the-ground observations. To do this we can use time-based analysis to measure the intensity of the tides and the landfall, and also use text analytics to help build these new observed hazard maps. While this is a much more speculative form of post-event analysis than those previously discussed, in an increasingly networked world it has significant potential to provide detailed real-time data for post-event situation analysis and management.

III. The Exposure Analysis Framework

In this section we describe our approach to designing of a framework that can make better use of existing exposure data to provide new ways of storing and analyzing this data. The framework provides a way of storing many types data, operating on this data, and analyzing the results from these operations all on a high-performance platform that can scale to handle very large exposure portfolios.

A. The Data Life Cycle

Currently, exposure data is often only used as input into pricing models. After being used for pricing, the exposure data is not used again. This process is shown in Figure 1.

With the Exposure Analysis Framework we hope to change this life cycle and ensure that we can tap into this rich data source. To do this we still use the exposure data for pricing, but after generation we move the data into a data warehouse where we can perform analytical operations on it. Not only can we import exposure data, but we can also store client, claims, event, and historical data inside

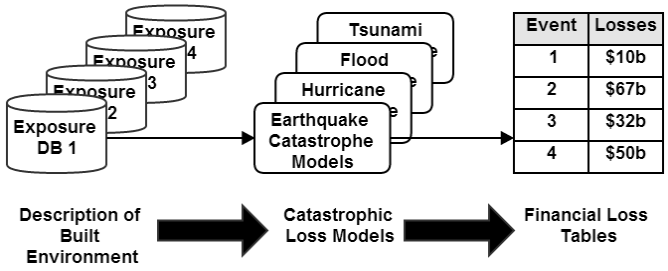

Figure 1. The current data life cycle for exposure data.

this warehouse. By importing this other data we can perform many new types of cross analysis, and visualize an exposure portfolio and risk in ways that were not possible previously. Figure 2 shows the proposed data life cycle.

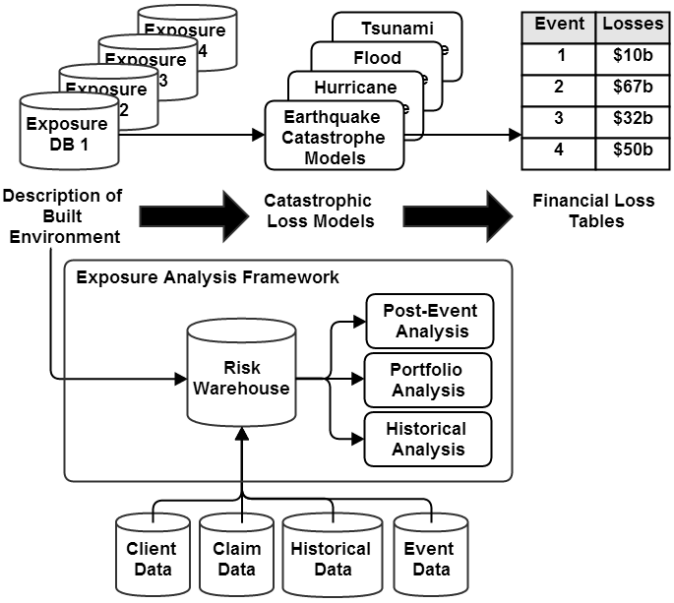

Figure 2. The proposed data life cycle for exposure data using the framework.

B. Operations

The analytical framework is designed around five fundamental types of operations, namely: Aggregation, Disaggregation, Geospatial, Loss Modeling, and Spatial Accumulation. These operations are central to exposure analysis and were identified through discussions between industry partners.

1) Aggregation Operations: The framework supports both standard and specialized aggregation operations. All aggregation operations group the data by some specified key or compound key and then perform some meaningful calculations on the data in order to reduce it to some generalized data value. Typical aggregation operations include minimum, maximum, average, total, and count operations.

Specialized aggregation operations perform actuarial and insurance-specific financial calculations.

2) Disaggregation Operations: Disaggregation operations take aggregated industry data and transform it into meaningful, finer-detailed exposure data. These operations are particularly useful when you have aggregated data and want to run it through a model. Since aggregated data is a coarse grained representation of data and models require data with finer details, disaggregation is used to transform this coarse data into finer detail. The disaggregation process is shown in Figure 3.

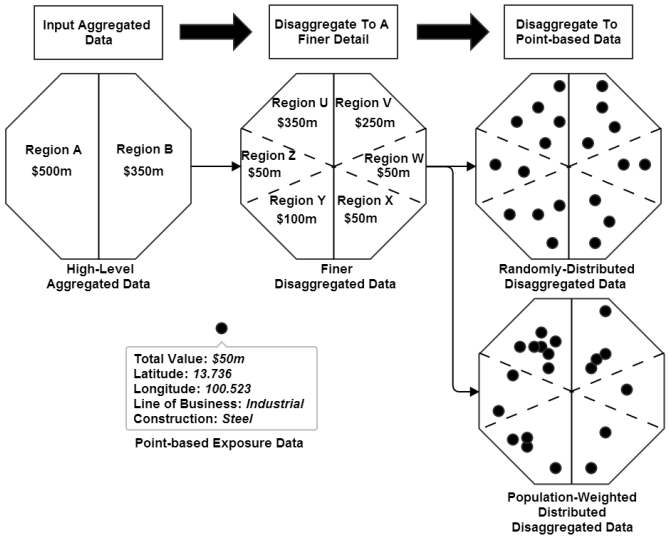

Figure 3. A graphical representation of how a disaggregation operation works.

The idea for the disaggregation is to transform this coarse, high-level data into geographical point-based exposure data. Disaggregation operations can transform data by in a variety of ways including distributing the data randomly within a specified region, or distributing the data based on population density.

3) Geospatial Operations: Geospatial operations provide a set of tools that can execute geographical queries and handle region data. With these operations you can perform point location queries in polygonal subdivisions, geocoding operations, and geometric operations on points and polygons. These operations also provide regional comparison tools, which provide a way to "redistrict" the regional boundaries of older, aggregated exposure data into newer regions in order for the data to be correctly represented in newer models. This process is shown in Figure 4. Redistricting is performed by disaggregating the data, placing the data points onto the old regional boundary map, either by assuming an even spread of points or by weighting the point data based on population, then modifying the boundaries to reflect the new changes.

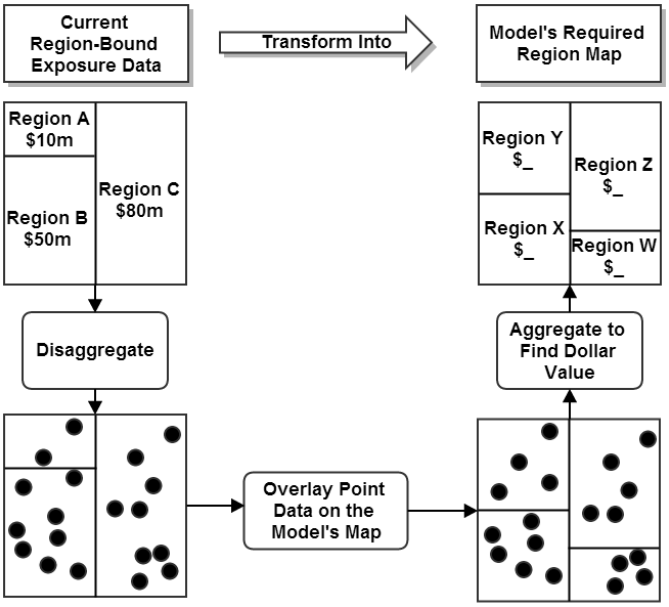

Figure 4. A graphical representation of how a boundary transformation operation works.

4) Loss Model Operations: Post-event loss modeling operations come in two flavors: single-event footprint loss models, and aggregate loss models.

Footprint loss models are used to compute the losses for a given single event. Event data is collected from a hazard map, which is used to calculate a mean damage ratio (MDR) for the event regions. Exposure data is then overlaid on top of the event regions and total losses are calculated by taking the total value of the exposure in the region and multiplying it by the MDR to find the total losses in the area. The result is a table of losses that can be filtered using a variety of keys.

Aggregate loss models are similar to footprint models, except instead of one event you have multiple events. The process still follows the same structure as the footprint model, however multiple hazard maps are used for each of the events. These models are used to illustrate the total combined losses of events and can also be filtered using a variety of keys.

5) Spatial Accumulation Operations: Spatial accumulation operations provide a method for identify regions of largest risk. These operations allow you to find a list of the largest non-overlapping risk or exposure concentrations within query circles of a given radius. Such circles represent exposure or risk accumulations and can provide companies with insight into the spatial distribution and/or clustering of their risks.

IV. The eXSight Framework - v0.1

Our current implementation of the Exposure Analysis Framework, called the eXsight Framework, is built on the

Java platform and has a MongoDB backend. MongoDB provides a robust suite of NoSQL operations that manage data handling and calculations for the framework.

A. Components

All of the framework's operations are implemented in Java classes and communicate with a MongoDB backend to manage data, execute operations, and handle results.

MongoDB is a versatile and scalable NoSQL database system that provides an alternative method of data storage to common SQL data stores [8]. Unlike regular SQL servers, where data is stored in a structured data schema, MongoDB stores data as a collection of documents containing various fields represented by a JSON-like structure[9]. Because of these differences in structure, traditional SQL tools and techniques will not work, and thus MongoDB provides a suite of robust tools and query engines that provides the same functionality that works with this new structure.

The eXsight framework utilizes MongoDB's geospatial indexes and query engine, Aggregation Pipeline, and MapReduce Framework to perform the various operations, while also utilizing the built-in NoSQL database functionality to handle data management [9].

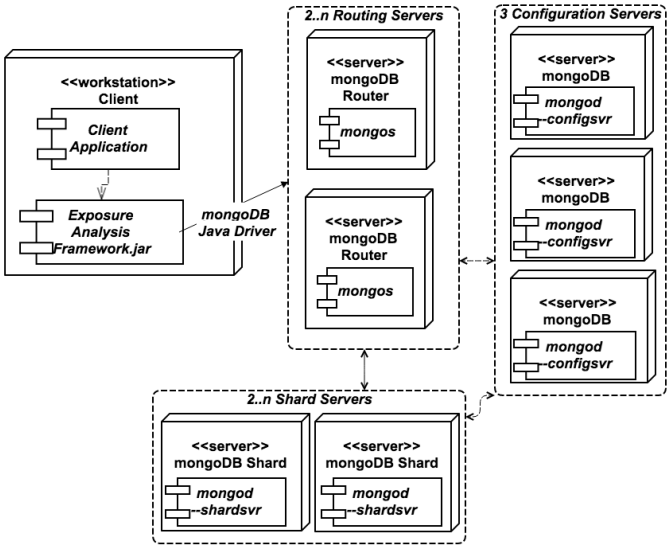

Figure 5. Deployment diagram depicting the recommended hardware architecture for the framework.

MongoDB can easily scale vertically by adding in more servers and horizontally with the use of sharding. Sharding is a form of horizontal scaling that allows data to be stored across multiple machines. As you can see in Figure 5, a sharded MongoDB backend consists of multiple servers. A sharded production database consists of two or more routing servers, exactly three configuration servers, and two or more shard servers.

The routing servers act as a middleman between the cluster and the application. They pass queries and operations on to the correct shards return the results to the client application. The configuration servers run are responsible for storing the cluster's metadata. This metadata keeps track of a map of the entire dataset, and knows which shard contains what data. This mapping is used to direct operations and queries from the routing servers to the appropriate shards. The shard servers are used to store a subset of the total data based on the type of data partition that is in use. MongoDB shards the data at the collection level based on a specified *shard key* which is present in every document in the collection [9].

Sharding is a crucial part of improving the performance and efficiency of the framework, however the current prototype framework only implements a single MongoDB backend instance. We will be extending on this in a future release.

To utilize the framework, a user will need to set up the MongoDB backend and call the framework's API methods to import data, perform operations, and handle results. The user can import data into the MongoDB backend and perform a selection of aggregation and geospatial operations by calling the appropriate methods.

B. Data Handling

The framework implements a generic data importer that lets the user define the overall schema of the data. The importer provides a way of ingesting exposure data without knowing its structure. To do this the user must describe their data either in Java objects or in an XML input file. The importer class takes as a parameter a configuration object, which is an instance of the class that defines the data schema. This schema has a list of provided data sources such as files, SQL tables, or NoSQL collections. Inside of each source you would describe the data through providing its location, datatype, and name. This is done for each element in order to build up a schema that the importer can parse to understand how to handle the data.

The framework also implements a way of exporting the results of operations by providing the user with different methods of handling and consuming results. The exporter allows the user to retrieve a Java string or list of the results, export data to a MongoDB collection, or even return the raw data.

C. Operations

In this section we describe how the core aggregation and geospatial operations are implemented using MongoDB's Aggregation Pipeline and geospatial indicies.

1) Aggregation Operations: MongoDB's Aggregation Pipeline is used to perform simpler aggregation functions. The pipeline works in a similar fashion to the UNIX pipe operator. The entire collection is passed through multiple operators to eventually reduce the collection to a single document containing the results [9].

2) Geospatial Operations: The eXsight Framework implements a set of operations that can execute geographical queries, and populate or correct administrative region data.

Users can find the administrative regions of a given point by providing a lat/long coordinate. For example, the point [28.418749, -81.581211], if queried, will return United States, Florida, and Orlando County as the country, state, and county of the point. This is done by using MongoDB's geospatial capabilities [9] to search through a collection of polygons representing a region and testing whether or not the given coordinate lies within the region.

The framework also allows users to geocode their data. Geocoding is used to map a set of points to their named region. The framework's geocoding operation takes this region query and runs through the entire exposure collection to populate the regional data for a given exposure point. This provides users with the ability to populate missing or correct inaccurate data.

V. Related Work

Currently Risk Management Solutions (RMS) offers a similar product to our framework called RMS(one). RMS(one) is a suite of tools on top of a high-performance backend which is accessible by a user interface and located in the cloud.

RMS(one) lets users ask advanced questions about their risk portfolio by visualizing model results and business metrics, automating and managing data importation, and viewing and managing user specified risk models. This is all through an intuitive interface running on top of a high-performance backend called the RMS Analytic Operating System (AOS)[10]. A portion of the RMS(one) backend utilizes MongoDB as a data store, managing and storing all of the client's data using MongoDB's flexible data model[11].

Our Exposure Analysis Framework has some similarities and differences to the RMS(one) product. Like RMS(one) the framework provides a suite of tools to help answer questions around a client's portfolio. We also utilize MongoDB as a main backend, however the current specification for our framework does not utilize cloud-hosting platforms. Unlike RMS(one), the framework does not have a user interface. Instead we provide an API that clients can build and application around to fit their needs.

Although both RMS(one) and the Exposure Analysis Framework has similarities, each has a few key differences which can allow both products to co-exist in the current market.

VI. Future Work

We are currently working to refine the design and implementation of the remaining core operations. Planned future work on the eXsight framework includes implementing the missing operations and optimization of operations. Since our goal is to have the framework operating on a sharded MongoDB cluster, we will also be migrating from a single MongoDB instance to a multi-node MongoDB backend.

We also plan a trial with industry partners to explore the practical applicability of the framework and to identify any critical new operations or functionality, input formats that need to be accepted, or output formats and data visualization options that need to be implemented.

References

[1] AON Benfield, Annual Global Climate and Catastrophe Report, Impact Forecasting 2013. [Online]. Available: http://thoughtleadership.aonbenfield.com/Documents/20140113_ab_if_annual_climate_catastrophe_report.pdf [Accessed: 13 Apr. 2014].

[2] P. Dunbar et al., "Tohoku Earthquake And Tsunami Data Available From The National Oceanic And Atmospheric Administration/National Geophysical Data Center," Geomatics, Natural Hazards and Risk, vol. 2, no. 4, Nov. 2011, pp. 305-323. [Online]. Available: Taylor and Francis Group2, doi: 10.1080/19475705.2011.632443 [Accessed: 30 Mar. 2014].

[3] Thai Meteorological Department, "Rainfall and severe flooding over Thailand in 2011," Nov. 2, 2011. [Online]. Available: http://www.tmd.go.th/en/event/flood_in_2011.pdf. [Last accessed: 31 Mar. 2014].

[4] "The World Bank Supports Thailand's Post-Floods Recovery Effort," The World Bank, Dec. 13, 2011. [Online]. Available: http://www.worldbank.org/en/news/feature/2011/12/13/world-bank-supports-thailands-post-floods-recovery-effort. [Last accessed: 31 Mar. 2014].

[5] "Hard Drive Prices Rise Due To Thai Floods," InformationWeek, Jan. 9, 2012. [Online]. Available: http://www.informationweek.com/data-protection/hard-drive-prices-rise-due-to-thai-floods/d/d-id/1102133. [Last accessed: 31 Mar. 2014].

[6] E. Blake et al., "Tropical Cyclone Report Hurricane Sandy (AL182012) 22 29 October 2012," Feb. 12, 2013. [Online]. Available: http://www.nhc.noaa.gov/data/tcr/AL182012_Sandy.pdf. [Last accessed: 31 Mar. 2014].

[7] D. Porter, "Hurricane Sandy Was Second-Costliest In U.S. History, Report Shows," Feb. 12, 2013. [Online]. Available: http://www.huffingtonpost.com/2013/02/12/hurricane-sandy-second-costliest_n_2669686.html. [Last accessed: 31 Mar. 2014].

[8] MongoDB Overview website: https://www.mongodb.com/mongodb-overview [Last accessed: 25 Mar. 2014].

[9] MongoDB Reference website: http://docs.mongodb.org [Last accessed: 20 May. 2014].

[10] RMS(one) Information website: http://www.rms.com/rms-one/rms-one#cloud [Last accessed: 1 April 2014].

[11] "RMS Revolutionizes Risk Management for Insurance Industry with Secure Platform Built on MongoDB," MongoDB, January 15, 2014. [Online]. [Last accessed: 1 April 2014].