AD-HOC EP CURVE CALCULATION USING MAP-REDUCE AND UI INTEGRATION WITH QLIKVIEW

Ashish Gupta, Abhishek Gupta
Big Data and Analytics, Data Solutions, AIG, Bangalore, India

ABSTRACT

Before Hurricane Hugo swept through Georgia and North and South Carolina in 1989, the insurance industry in the U.S. had never suffered a loss of more than $1 billion from a single disaster. Since then, numerous catastrophes have exceeded that figure. Hurricane Andrew in 1992 caused $15.5 billion in insured losses in southern Florida and Louisiana. This trend is growing and risk analysis is a key aspect for these loses. In this paper, we explore the design of a framework for portfolio risk modeling for AIG insured locations. The proposed model of the framework integrates the portfolio risk modelling methodology with user interface Qlikview which eases the process of ad-hoc analysis and can provide the end result with-in a day. The goal was to provide the end user the capability to do ad-hoc analysis with different set of dimensions and without writing the complex SQL queries. This framework will be useful for Insurers and risk managers to assess the risk in a portfolio of exposures. This will also be helpful for underwriters to decide the reinsurance.

KEYWORDS: *Portfolio risk, Map Reduce, Hadoop, Qlikview*

I. INTRODUCTION

Residential and commercial development along coastlines and areas that are prone to earthquakes and floods suggest that future insured losses will only grow — a trend that emphasizes, as never before, the need to assess and manage risk on both a national and a global scale.

Exceedance probability (EP) Curves [6] gets people to think about risks on a broader level. With the EP Curve an insurer will have the information about critical aspects such as likely damage, probability that the loss will be greater than $X million or $Y million.

During the portfolio loss modeling [6], Event Loss Tables (ELT) [6] are used which are building blocks of EP Curve. Since ELTs are at transaction level and not stored for reuse, this end-to-end portfolio roll-up process takes up to 1-3 months. Ad-hoc analysis is difficult and is time-consuming.

To calculate the financial impact of natural catastrophe, insurance industries use catastrophic modeling (CAT modelling) [5]. Portfolio loss modeling may consists of reinsurance contracts covering millions of individually insured locations. To calculate annual portfolio risk, each portfolio must be evaluated in up to a million of simulation trials, and each trials consists possibility of a lot of catastrophic events, such as floods, wild fire, earthquakes. CAT modeling is one way of portfolio loss modeling but as of now this process is time consuming and supports only predefined dimensions.

This paper proposed a framework which facilitates to answering rich set of ad-hoc queries for generating EP (Exceedance probability) Curves for different dimensions and helps building portfolio loss modeling. Key feature of proposed framework is that it is designed for the users who are good in understanding CAT modeling but having little or no skills of programming and SQL language. The user selects the dimensions of interest and submits his selection just on a click of button. Frameworks take care of building Event Loss Tables (ELTs) and calculate EP Curves and provide results back to users. The complexity of huge data and calculations of aggregated ELTs and EP Curves are hidden from user.

We implemented a prototype system called AEPIQ (Ad-hoc EP curve calculation using map-reduce and UI (User Interface) Integration with Qlikview [4]). This system allows user to take advantage of Hadoop [1] streaming jobs for faster processing and Apache Hive [2] to store large scale data. To save user from writing complex SQL queries for ELT generation and to make ad-hoc queries possible, we integrate Qlikview with Apache Tomcat [3].We describe the design and implementation of AEPIQ and present experimental results.

II. RELATED WORK

In this section we will focus on the related work that has been done previously by several researchers. Portfolio loss modelling for risk analysis has been a focused attention. Much research is currently being conducted in aggregate risk analysis and portfolio risk analysis using EP curves [6]. This analytical technique has become a choice of many modern insurance companies. For instance, a reinsurance company holds a portfolio of programs to insure primary insurance companies to cover large scale losses. Each program contains data that describes the buildings to be insured, the modelled risk (ELTs), and layer contracts. The modelled risk is represented by an ELT [6]. Every row of the table covers different building. The attributes of the table can be building's location, construction details, insurance coverage, and replacement value. This table lists the expected loss that would occur to the buildings if any catastrophe occurs. Layers are defined as per the company's contract with primary insurance companies.

However the process of risk analysis is usually a tedious process and time consuming, which involves analyzing risk data, writing complex SQL queries, generating ETLs and EP Curves. With the help Qlikview [4], not only is this possible with a click of a button, but also efficient.

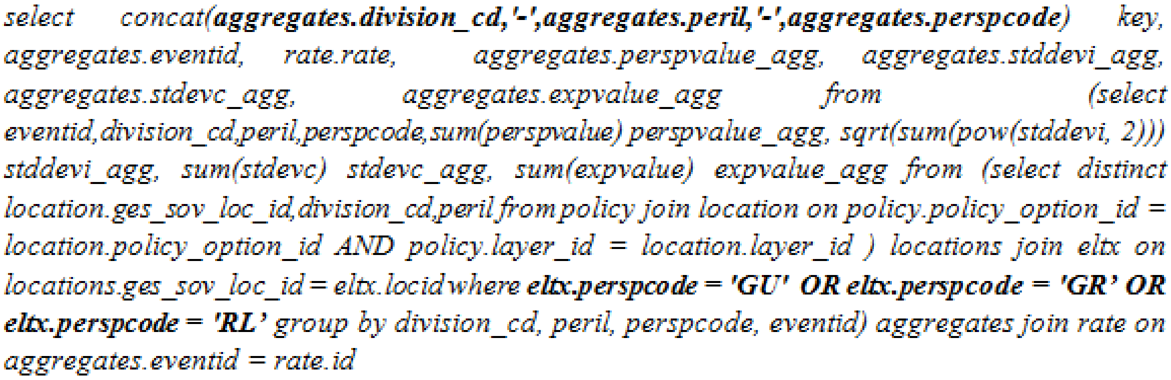
III. METHODOLOGY

To build this model we are already provided predefined tables for different dimensions. These tables have data for policy, location and event ids for different layers. These tables are saved into Hive.

In order to answer ad hoc query efficiently AEPIQ pull policy table, having different dimensions, into Qlikview. Now user can select dimensions for which he wants to build the EP Curves.

3.1 Select Dimensions & Generate Queries

Selected dimensions are passed to a web-application which is deployed on the Tomcat. Based on selected dimensions, a Hive query will be built, which takes advantage of the Map-Reduce programming model to calculate the aggregated ELT table. In the normal process there are four different complex SQL queries to get the ELT and that too for a specific value of dimensions. After the query, result will be saved on HDFS [1].



```
select concat(aggregates.division_cd,'-',aggregates.peril,'-',aggregates.perspcode) key,
aggregates.eventid, rate.rate, aggregates.perspvalue_agg, aggregates.stddevi_agg,
aggregates.stdevc_agg, aggregates.expvalue_agg from (select
eventid,division_cd,peril,perspcode,sum(perspvalue) perspvalue_agg, sqrt(sum(pow(stddevi, 2)))
stddevi_agg, sum(stdevc) stdevc_agg, sum(expvalue) expvalue_agg from (select distinct
location.ges_sov_loc_id,division_cd,peril from policy join location on policy.policy_option_id =
location.policy_option_id AND policy.layer_id = location.layer_id ) locations join eltx on
locations.ges_sov_loc_id = eltx.locid where eltx.perspcode = 'GU' OR eltx.perspcode = 'GR' OR
eltx.perspcode = 'RL' group by division_cd, peril, perspcode, eventid) aggregates join rate on
aggregates.eventid = rate.id
```

Figure 1: SQL Query

3.2 EP Curves Calculation

Now from the web-application, a different thread will trigger the EP Curves calculation. This will call a mapper program which will create the key value pair. Key would be selected dimension's concrete value and values would be corresponding ELT table. These key value pair will be passed to the reducer program which will calculate the EP curves for different keys.

3.3 Results Fetched into HIVE

While this calculation is running user will be return the job id of the Map-Reduce [1] job. User can view the status of the job from this id on the Qlikview UI. After completion of the job, results will be saved into Hive table Ep Curves.

3.4 Integration with Qlikview

Now taking a different sheet from the Qlikview UI, user can pass the concrete values of the dimensions for which he is interested to view EP curve. This will be passed to web application which will trigger the Hive Select query on EP Curves table and will return the result to web application, which will pass on the response to Qlikview and it generates a table and graph.
Throughout the process, there is no need to write any query or program for calculations.

IV. ARCHITECTURE

Steps involved in the architecture are mentioned below:-

- Users selected dimensions passed to the Tomcat using VB-script [3] inside Qlikview.
- Parameters are passed to Java API which constructs Ad-Hoc Hive query. Hive query is executed on the Apache Hive.
- Resulted ELT tables for different combinations of selected dimensions are saved on HDFS.
- A new Java thread triggers the streaming map reduce job on Apache Hadoop cluster.
- Job ID of the launched job is returned to web application.
- Qlikview receives this job id as part of submitted query.
- Map-Reduce streaming job result is saved in Hive EP-curve table.
- User queries for a particular dimension value.
- Hive Select query result goes back to web application.
- Web application sends response back to Qlikview.

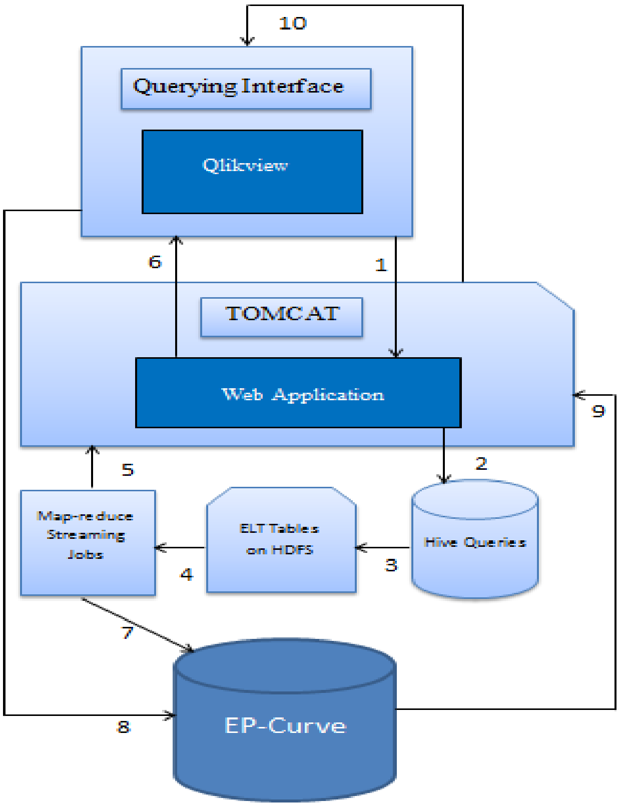


Figure 2: High Level Architecture Design

V. Results

Let’s say user wants to see EP curves for Division and Peril dimensions. He selects the check boxes in Qlikview UI corresponding to these dimensions and submits the query. In the program following query will be built automatically at back end:

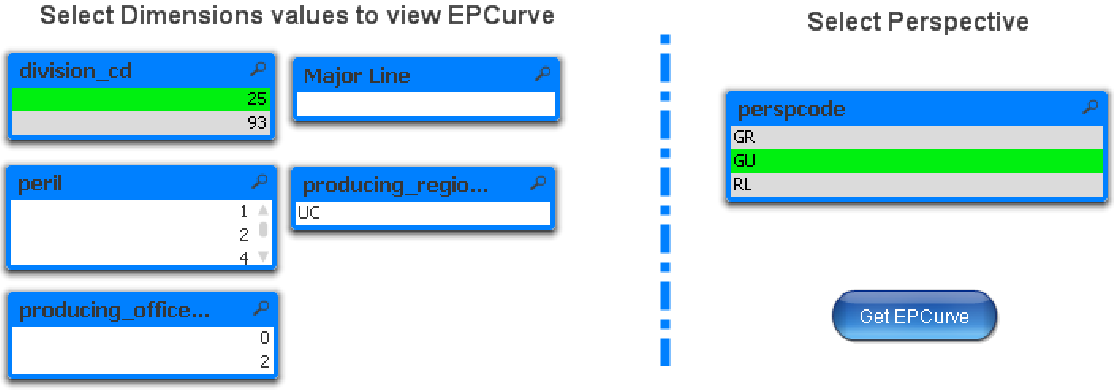
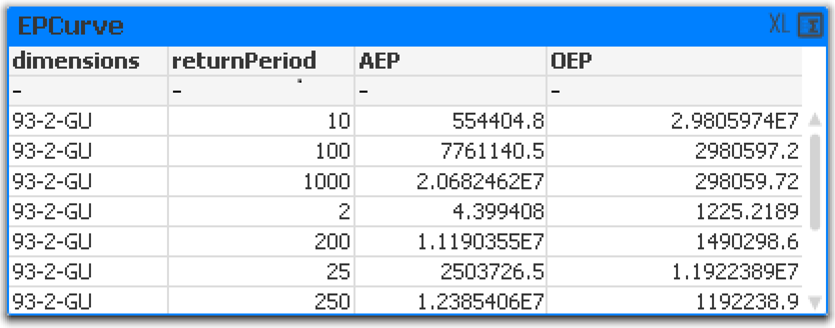


Figure 3: Selecting Dimensions

So program takes care for the query building part as shown in figure 1 and sends this query to Apache Hive. Hive executes this query and generates ELT tables for different combinations of dimensions. Examples of the results are:

93-1-GU, 2847001, 1.0E-10, 4.245479583740234, 36.086578369140625, 6.36821932, 298342.59375

93-2-GU, 2847004, 1.0E-10, 29937. 66015625, 173703.859375, 30653.623046875, 6947418.5



dimensions	returnPeriod	AEP	OEP
-	-	-	-
93-2-GU	10	554404.8	2.9805974E7
93-2-GU	100	7761140.5	2980597.2
93-2-GU	1000	2.0682462E7	298059.72
93-2-GU	2	4.399408	1225.2189
93-2-GU	200	1.1190355E7	1490298.6
93-2-GU	25	2503726.5	1.1922389E7
93-2-GU	250	1.2385406E7	1192238.9

Figure 4: Result

So let’s assume we have 2 divisions, 3 peril in the table. In one go, query will generate 18 different combination of ELT tables (3 different perspective GU, GR, RL for every combination).

Now for these different combinations Mapper will separate the ELT’s and simultaneously 18 reducer programs will run in parallel. So one query will generate 18 different EP Curves.

User will now query with the values of dimensions to get the EP curves (For example division = x and peril =y and perspective =GU).

VI. Conclusion

In summary our experiments shows that AEPIQ is capable of handling ad-hoc portfolio loss modeling queries on industry size data sets in matter of minutes. It does not require user to have SQL or programming experience and can be used interactively using Qlikview interface. User just needs to select the dimensions in Qlikview. The generated complex queries are fetched into our framework and results are displayed clearly in Qlikview.

A user using AEPIQ, for 2200,000 records, processing on 20 node clusters took 9 minutes to generate 6 combinations of EP curves and which are articulated in Qlikview.

VII. FUTURE WORK

Our aim is to enhance this system so that it can be customized based on user's need. We are also targeting to make this process generalized for other Business Intelligence tools other than Qlikview. In future we will try to include Apache spark to further improve the processing time of queries.

REFERENCES

[1]. Apache Hadoop. http://hadoop.apache.org/.
[2]. Apache Hive. http://hive.apache.org/.
[3]. Apache Tomcat. http://tomcat.apache.org/
[4]. http://www.qlik.com/
[5]. R. R. Anderson and W. Dong, "Pricing Catastrophe Reinsurance with Reinstatement Provisions using a Catastrophe Model," Casualty Actuarial Society Forum, summer 1998, pp. 303-322.
[6]. A. Rau-Chaplin, B. Varghese, D. Wilson, Z. Yao, and N.Zeh, "QuPARA: Query-Driven Large-Scale Portfolio Aggregate Risk Analysis on Map Reduce," Vol. 1, 16 Aug 2013.

AUTHORS

Ashish Gupta is working as assistant manager in Big Data & Analytics Team at AIG Data Solutions India. He has pursued his under graduation in Information Technology from Kamla Nehru Institute Of Technology, Sultanpur. He has been working in the field of software development for the last 10 years.

Abhishek Gupta is currently working as Programmer Analyst in Big Data & Analytics Team at AIG Data Solutions India. He has pursued his B Tech. in Information Technology from VIT University, Vellore. His interest lies in Data modelling, Database concepts, Hadoop, Hive, Map reduce.