Cloud Analytics Proof of Concept | Briefing Paper

As in so many other areas, cloud computing has the potential to provide disruptive change to business intelligence. Leveraging emerging open-source technologies such as Hadoop, and renting space on servers, can potentially provide powerful data mining at a fraction of the cost of established technologies. To better understand this emerging field, Cable & Wireless Communications in partnership with DigitalRoute set out on a proof of concept in early 2011. This paper outlines the concept being tested, the results and some of the possible impacts this will have on the telecommunications business.

1. Introduction

Examples of the power of extracting insight from business data abound, such as Tesco mining their store-card information to know what combination of items sell. Extracting this value, however, is not always straightforward, as traditional data-warehouse technologies have been expensive and typically required large hardware and software investments. Moreover, depending on the type of information, storing and accessing long-term historic data can mean keeping very large amounts of data on-line in databases, with various challenges around costs, infrastructure, operations and performance (even if Moore's law of increasing computing performance makes it a diminishing problem).

But, as in so many other areas, Cloud computing and open-source technologies provide a radical change and create an opportunity to carry out analytics at dramatically lower cost. Compared to buying servers in the traditional way, our test cost only 0.0015%[1] or thousandths of a percent of the equivalent hardware purchase. By literally renting computers and storage only when one needs it and combining it with open-source technologies, analysing large volumes of data is suddenly within reach of any organisation. Not only does the price change fundamentally, the capability to add additional computing resources with some simple configuration (elasticity) means that such tests can quite easily tap in to hundreds of servers. Five years ago if you asked the average IT organisation how long it would take to deliver 100 servers, the organisation might have struggled.

To demonstrate how this could be done we ran a proof of concept (POC), using a year of call-data from our networks to extract roaming habits.

The second important aspect we wanted to demonstrate was how such analysis can be done on 'tokenised' data, which does not pose the same issues from a sensitivity and security perspective. With increasing occurrences of where systems have been compromised and security breached, the technology industry is likely to have to rethink how it manages some of its data.

How important such security is, became obvious in the POC when, just for verification purposes, we looked at known colleagues' and friends' data and suddenly knew to which countries they had travelled, who they had called and so on. With these technologies communication service providers could easily peek deep into the network and track both internet and other usage behaviours. No wonder there are concerns about privacy.

[1] Not a strict "apples for apples" comparison, but just comparing an investment of $30,000 for hardware and software vs. our average cost to run a single 10 million record query on Amazon of $0.44.

2. THE TEST

To demonstrate the concept we used MediationZone® from DigitalRoute to retrieve all the roaming data for post-paid customers in one of our regions. This did not represent a particularly large data-set, but we also created alternative sets with 750,000,000 records to test the concepts and get a feel for performance and costs.

Before we shipped the data to the Amazon storage buckets (we used S3), we pre-processed all the records in MediationZone. Through a single workflow, we replaced all phone-numbers with tokens (see below) and also filtered, formatted and compressed the data so that the original size was reduced by about 58 times[2]. This was done by removing all fields we were not interested in and by using compression (gzip in this case), which on average achieved 90% compression rates.

We then used the MediationZone Amazon connector to upload the data directly to Amazon S3 storage. From here we used Amazon's Elastic MapReduce, which implements Hadoop to carry out the analysis. One of the beauties of Amazon Web Services (AWS) is that it comes with Hadoop pre-configured, so starting a job even with hundreds of servers takes just a few clicks.

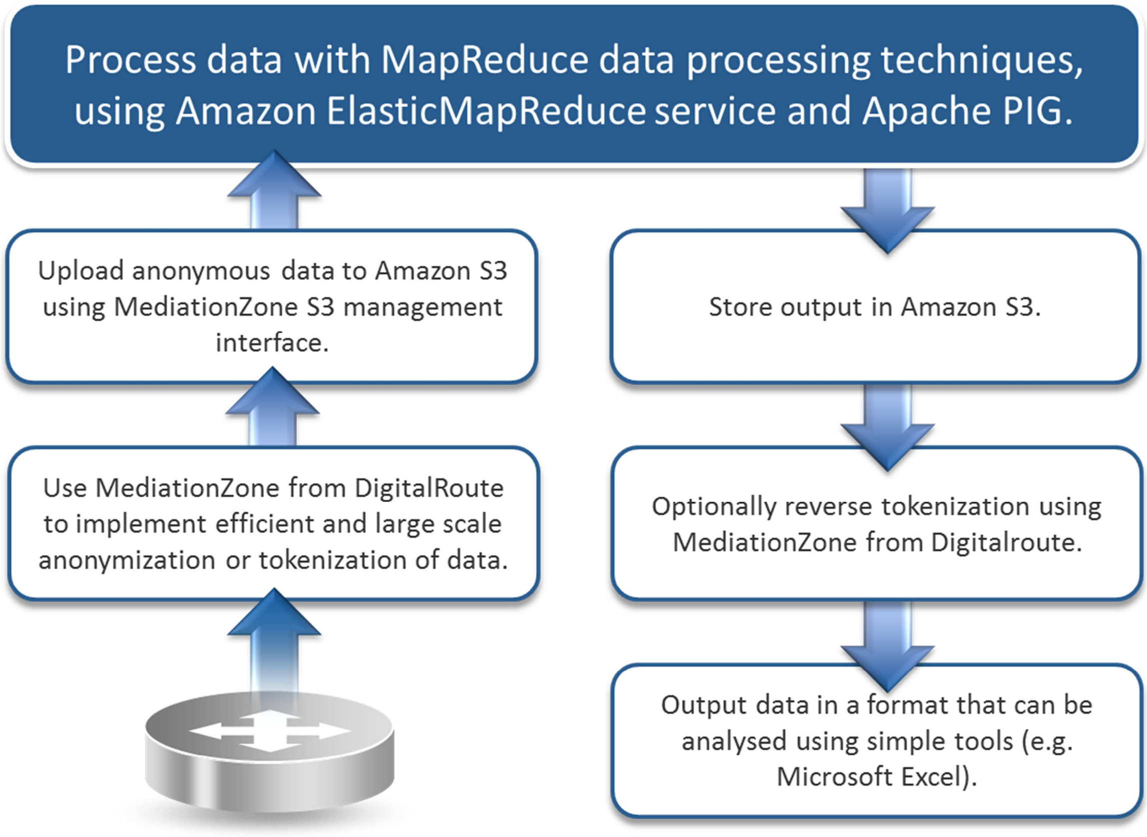
Process data with MapReduce data processing techniques, using Amazon ElasticMapReduce service and Apache PIG.
Upload anonymous data to Amazon S3 using MediationZone S3 management interface.
Use MediationZone from DigitalRoute to implement efficient and large scale anonymization or tokenization of data.
Store output in Amazon S3.
Optionally reverse tokenization using MediationZone from Digitalroute.
Output data in a format that can be analysed using simple tools (e.g. Microsoft Excel).

Figure 1: Overview of the processing flow carried out as part of the POC.

[2] The benefit of compressing the data is the reduced costs for transfer and storage and the reduced time it takes to upload the data.

We configured Hadoop (by using Apache Pig and Pig-Latin) to answer five questions:

- When roaming in a country, which countries are most called and texted? (Grouping the data by user, country roaming in and the target countries for calls, text and data).
- Which accounts and services travelled the most abroad, how many calls did they make and texts did they send? (By summarising the data).
- To which countries did they travel? (By counting the unique days on which they made or received calls, text or data in each unique country).
- To which countries did they call (by summarising the destinations).
- Which account & services travelled the most, calculated as one separate trip if there were no roaming calls for 72 hours or three days in a row (by sorting the data and looking for gaps that lasted more than 72 hours).

The last question was designed to create a fairly difficult query, which meant that data could not just independently be summarised, but needed to be sorted and then looked at in context.

The output from each query was stored in a comma separated file which could then easily be manipulated for further analysis. Finally we could download the results from AWS, and 'de-tokenise' the data should we for example want to contact a particular account.

In order to test the scaling of data, we also ran a number of extrapolated tests, where we copied the roaming data to generate larger volumes and processed this to see how Hadoop performed.

3. Security and Tokenisation

A key concern these days is security. Often for businesses and IT professionals, there is a feeling that when the servers sit in their own premises the security is better than in the cloud. However, this sense of security is often not justified. The risk profile does not necessarily change, whether a server is in the cloud or in a local office space, as long as the access to any cloud storage is as equally controlled as the office connectivity (through secure tunnels and restricted IP ranges for access).

At the same time, a cloud-provider is frequently certified and has implemented security-processes that far outstrip the average IT department. After all, for the cloud business security is core to their survival and winning new contracts.

Even so, the type of data that is processed by analytics is often sensitive, as in this example, where the data contains a person's phone records by which one could determine their travel patterns, who they call and who calls them. This data should be protected, more from a data-privacy perspective than cloud per-se and this should be considered even in internal installations.

Therefore a key element of the POC was to demonstrate that analytics can be carried out on data that has been made much less sensitive by replacing key personal information with randomly generated 'tokens'. The purpose of this "tokenisation" is to ensure that a person cannot be identified by the data alone (no conversation algorithm is used, but true random sequences to ensure there is no way to reverse the calculation). At the same time, the tokenisation is done such that the relationships between numbers are maintained. For example, the fact that A called B and B then called A can still be determined, but who A and B are, is not visible.

The tokenisation process generates a key/value repository, which is maintained in a locked down environment locally, such that the encryption can only be reversed using this information. This tokenisation makes the data far less sensitive from a privacy perspective without limiting the possible analytics.

4. Results

The first and perhaps most eye-catching result is the cost of processing. After various test-runs, uploads and so forth, the total bill for hardware, including costs to transfer data, storage and processing, was $18.18. The cost to run a single small 10m records job for analysis came to $0.44, including servers, storage and data-transfer. For all our tests, including running a couple of tests with close to a billion records, the bill still ended up less than a hundred dollars (yes, we paid by credit-card).

Once the data-feed was established and formats were proven, the estimated time to write a query is very comparable to something in SQL.

We did not spend much time investigating the performance of Hadoop, as we did not really have sufficient data to really scale it (Hadoop emerged from the work that Google did on MapReduce and is used by web2.0 companies to process logs and other data from their massive user-communities by dividing up the work in many parallel streams). In our case, we ran a couple of tests with one to ten instances to get a feel for the balance between performance and costs.

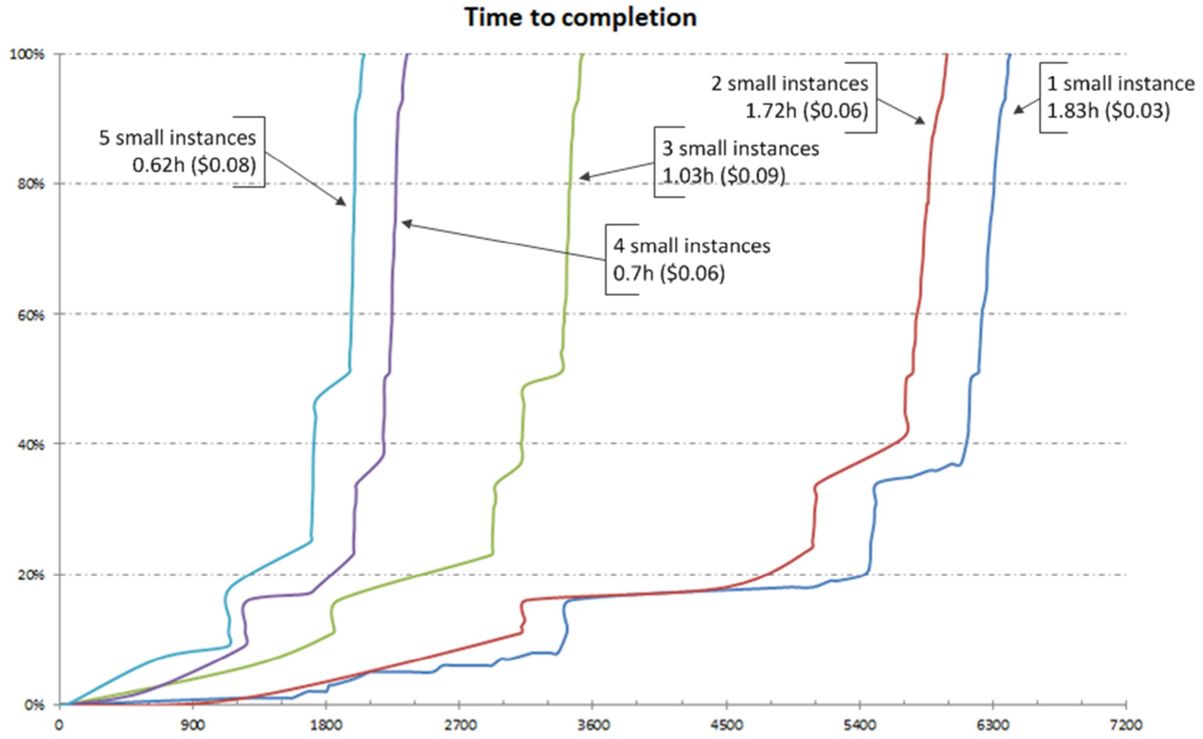

Figure 2: Sample of how Hadoop scaled in time and instances.

The results allowed us to run a single small instance, which completed in 1.8 hours. If we instead run five instances, it completed in 37 minutes, so though we did not see a direct linear increase, that tasks completed more quickly. The cost for the servers in this case was $0.20 for the single instance and $0.50 for five. Hadoop is known to scale very well, so this is more likely to be a question of tuning.

We also ran some larger tests. Some of our larger tests completed in approximately 33 hours, with 10 instances.

All this should be compared with typical hardware implementation and running costs. We would easily spend [$]25,000 and more on a small-ish business intelligence server, once hardware, disk software licenses as well as running resources have been included. And of course most hardware sits idle most of the time (the US Government estimated that their average utilisation of hardware was 6%), which creates unnecessary energy wastage.

Another concern frequently raised is connectivity. In our case, the upload to AWS was a single stream, which we did not spend a lot of time tuning, as it completed in 16 minutes. The downloads were even faster.

5. BUSINESS IMPACT AND OPPORTUNITIES

What does this mean for businesses? The first and most important conclusion is that the availability of these cost-efficient technologies can enable even the smallest company access to large data mining and intelligence.

For telecommunications companies, which typically have regulatory demands such as storing and retrieving large amounts of data, moving to ‘nosql’ technologies (as these are referred) can offer a very compelling cost alternative.

The tokenisation of data also enables a freedom to use these technologies more freely. For example companies that are highly regulated, such as telecoms, finance and insurance can use tokenised data to still see trends, find big spenders and analyse behaviours. Should they at any given point need to reverse the data, it can be de-tokenised by accessing the local and secure key/value pairs.

In our case, Cable & Wireless Communications will look at leveraging these technologies to augment our business intelligence systems to store years of detailed data that would otherwise not be possible, and thereby enable the business to go back and re-create summaries, even for data that is many years old. Moreover, with the explosion of mobile data, we can chose to store an increased amount of the detail that would otherwise just have been thrown away, to learn more about our customers’ habits and what they might be interested in. We will do this by leveraging technologies such as Hive to expose the capabilities within the regular reporting environment.

Having said this, we do not foresee that these technologies will replace daily operational reporting (just yet) and some other specialised data analysis technologies. We imagine that this type of set up is particularly attractive to support ‘what if’ type of questions and perhaps more importantly a capability to store and query detailed historic data for much longer periods than is possible today.

6. Conclusion

With these new technologies all companies are getting access to huge processing power at an unprecedented scale and radically different cost-levels. For telecommunications this represents an attractive opportunity to reduce the cost and gain much better insight into what customers are doing on networks, something that is becoming more urgent with the large volumes of data generated from the new mobile devices.

However, with so much detail of what our customers are doing at our hands, data stewardship, security and privacy become urgent topics. It is a fine balance between offering relevant packages and breaching someone's trust. The tokenisation helps in keeping data anonymous and making it much less sensitive, but particularly around the marketing aspects, great care and integrity is required. This is certainly highlighted by all the recent press around both iPhone and Android and the data it generates and stores.

As so often with new technology, the hardest question of all might actually be somewhat surprising: What question do we want to answer from all this data?

About Cable & Wireless Communications

Cable & Wireless Communications is a global, full-service communications business. We operate leading telecommunications businesses through four regional units – the Caribbean, Panama, Macau and Monaco & Islands. Our services include mobile, broadband and domestic and international fixed line services in most of our markets as well as pay-TV, data centre and hosting, carrier and managed service solutions. Our operations are focused on providing our customers – consumers, businesses, governments – with world-class service. We are the market leader in most products we offer and territories we serve. For more information visit www.cwc.com.

About DigitalRoute

DigitalRoute® is a Swedish independent software vendor delivering market leading mediation and integration solutions to the global telecommunications and data communications industry. DigitalRoute simplifies service providers' data infrastructure, centralizing integration of IT- and communication networks. DigitalRoute technology for mediation-, data integration and policy control is deployed at 220 customers worldwide. For more information, please visit www.digitalroute.com.

7. Appendix: Technology References

The technologies we used in the Proof of Concepts were:

Amazon Web Services are a set of services offered by Amazon, where companies through the internet can rent computing resources such as servers and storage. AWS is today by far the largest provider of infrastructure-as-a-service or ‘public computing clouds’.

MapReduce is a framework for processing huge data sets on certain kinds of distributable problems using a large number of computers. The advantage of MapReduce is that it allows for distributed processing of the map and reduction operations.

Apache Hadoop is a software framework, inspired by Google’s MapReduce work that supports data-intensive distributed applications.

Apache Pig is a platform consisting of a high-level language (***Pig Latin***) for expressing data analysis programs, and a compiler that produces sequences of Map-Reduce programs suitable for running on Hadoop.

Amazon Elastic MapReduce is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon EC2 and S3. Specifically, Elastic MapReduce workflows can run analyses expressed using Apache Pig to process data stored in S3.

MediationZone is a mediation platform that supports any data exchange between any systems, in both online (streaming) and offline (file based) transactional modes.

MediationZone provides high-level graphical configuration using workflows to model data process flows. Workflows are based on software agents that are linked into mediation tasks of virtually any complexity. A powerful drag-and-drop management user interface covers all aspects of workflow life-cycle management.

MediationZone is based on a modular architecture, where functionality and processing capacity can be added over time. It delivers an execution environment for mediation applications using standard software components. There is a separation between the runtime platform, off-the-shelf functionality and configuration. This enables independent evolution of computing infrastructure, standard software and customer configuration as requirements change. In addition, MediationZone separates the concept of protocols, from physical/logical data formats and for whatever business logic is necessary for managing an information flow.