Reducing Risk in KYC (Know Your Customer) for large Indian banks using Big Data Analytics

Anuraj Soni
COO – Magic Software Pvt. Ltd.
9th Floor, Tower C, Tech Boulevard
Sector 27, Noida, UP – 201301, India

Reena Duggal
Technical Manager – Headstrong Inc.
D-4, Sector 59
Noida, UP – 201307, India

ABSTRACT
KYC (Know Your Customer) is becoming a critical gatekeeper process for financial institutions, the world over, to safeguard against financial frauds, terrorist funding and money laundering. It involves collecting basic identity & address information about the customer. Regulatory agencies have been coming down heavily on defaulting organizations thereby forcing many of them to invest in state of the art financial transaction surveillance systems. One of the biggest challenges that the industry faces while it steps up monitoring is the sheer size of the data, speed of generation of this data and complexity arising out of multiple & non-standard formats.

World continues to generate data at an unprecedented pace. This generation calls it Big Data. Traditional data warehousing techniques, that have stood the test of time, have lately failed to deal with volume, velocity & variety of data (Big Data). This is where Big Data Analytics has emerged victorious. Rather than relying on structured data techniques, Big Data analytics attempts to rely on basic techniques like pattern matching, divide & conquer & decentralized processing to solve real life problems. Though this technology is still new but it has already shown signs of maturity.

This paper attempts to study Know Your Customer process, articulate the challenges involved and highlights the short-comings that the systems today have in effectively implementing KYC guidelines(especially in large Indian banks). It then, using real life examples, presents a credible solution using Big Data Analytic techniques like Fuzzy Matching & MapReduce. Authors are confident that the framework of the solution that has been provided can lead to a working prototype in a short span of time.

General Terms
Big data analytics, Large Indian Bank, Know Your Customer

Keywords
KYC, Aadhaar, Unique Identity Proof, Fuzzy matching, Identity theft, MapReduce

1. INTRODUCTION
The approach for this paper is to first describe both Big Data Analytics & Know Your Customer (KYC) concepts in detail. It will then focus on the risks that remain unaddressed in KYC specifically for the Indian markets. Having clearly described the problem it will then be followed up with a solution framework using Big Data Analytics.

2. WHAT IS "BIG DATA ANALYTICS"?
"Big Data" referring to the massive and exponentially growing amounts of data that accumulate within many companies, has become one of the buzzwords in IT since the last couple of years. Every day the world create 2.5 quintillion bytes of data. In fact, 90 percent of the data in the world today has been created in the last two years alone. This data comes from a wide variety of sources: social network profiles; web site tracking information; server logs; sensor data; digital pictures; audio and videos; purchase transaction records and cell phone GPS(Global Positioning System) signals; document scans; financial market data; banking transactions - much of it generated in real time and in a very large scale. This is Big Data.

The "Big" part of big data may be defined in terms of the Three Vs: volume, velocity and variety.

1. Volume – the quantity of data relative to the ability to store and manage it

2. Velocity – the speed of calculation needed to query the data relative to the rate of change of the data

3. Variety – a measure of the number of different formats the data exists in (e.g. text, audio, video, logs etc.)[1]

If any of these Vs is low, then it may be more efficient to analyze the data using traditional BI (Business Intelligence) methods. However, if volume and required velocity are high, then big data techniques and technologies become more efficient and economical.

The word "Big Data analytics" means the systematic discovery of meaningful patterns and unknown correlations in large amounts of data of a variety of types (Big data) to support decision-making. It comprises tools and processes that help analyzing large amounts of organizational data.

Using advanced analytics techniques such as predictive analytics, data mining, statistics, and natural language processing, businesses can study big data to understand the current state of the business and track evolving aspects such as customer behavior. The technologies associated with big data analytics include NoSQL databases, Hadoop and MapReduce, massively parallel processing databases, in-memory database, search-based applications, data-mining grids, distributed file systems, distributed databases and cloud. These technologies form the core of an open source software framework that supports the processing of large data sets across clustered systems. [2]

Gartner expects the market for Big Data and analytics to generate $3.7 Trillion in products and services and generate 4.4 million new jobs by 2015.

3. WHAT IS "KNOW YOUR CUSTOMER (KYC)"?
The key to survival in today's financial services market can be summed up as: "Better know your customer." The

identification of a customer is a very critical process in KYC with a view to protect the customer interests by preventing from fraudsters who may use the name, address and forge signature to undertake illegal business activities, encashment of stolen drafts, cheques, etc. This also helps to safeguard banks from being unwittingly used for the transfer of funds derived from criminal activity or for financing terrorism. Identification of customers also helps in controlling financial frauds, identify money laundering and suspicious activities, and for scrutiny / monitoring of large value cash transactions.

In India, in order to prevent these issues, the Reserve Bank of India (RBI) had directed all banks and financial institutions to put in place a policy framework to know their customers before opening any account. This involves verifying customers' identity and address by asking them to submit documents that are accepted as relevant proof.

Mandatory details required under KYC norms are proof of identity and proof of residence. Passport, Voter's ID card, Permanent Account Number (PAN) card or driving license are accepted as proof of identity, and proof of residence can be a ration card, an electricity or telephone bill or a letter from the employer or any recognized public authority certifying the address, in addition to proof of identity being used as residence proof in case they carry address. More recently Aadhaar card (Unique Identification Number to all Indian citizens given by Unique Identification Authority of India (UIDAI)) is being used as a valid KYC document as both proof of identity and proof of address. Recent advancements have brought about eKYC (Electronic–Know Your Customer) using Aadhaar where only biometrics are provided and identity & address is verified online.

Some banks may even ask for verification by an existing account holder. Though the standard documents which are accepted as proof of identity and residence remain the same across various banks, some deviations are permitted, which differ from bank to bank. Similarly most high value financial transactions require customers to disclose their PAN numbers.

4. WHAT ARE THE RISKS INVOLVED IN KYC?

KYC norms set by regulatory agencies seem quite straight forward and if followed, can avoid most risk and fraud, provided the following three conditions are met:

4.1 Unique National Level Identifier

A unique national level identifier needs to exist which must be mandated to be collected as an identity proof for every financial and non-financial transaction. Once this gets implemented, the only risk involved is if someone were to obtain multiple identities using fraudulent means. Even this risk is substantially reduced if biometrics can be obtained because then de-duplication can easily detect an individual trying to obtain multiple identities [3]. For example, Social security number in USA is one such national level identity which needs to be provided in every financial and even for most non-financial transactions. However since it does not use biometrics and hence it leaves a probability for fraud. India's Aadhaar has attempted to even plug this hole by using biometrics as a way to check such fraud.

In India, even with Aadhaar, there are two challenges that need to be surmounted

1) Every citizen must get Aadhaar Card. Less than half the Indian population has Aadhaar card today. [4]

2) Aadhaar is mandated as mandatory identifier to be provided for every financial transaction.

Note: Even after 8 years of rollout of PAN card (initiated in Jan 2005), India hasn't yet been able to either provide PAN card to each Indian citizen or make it mandatory for each financial transaction. [5]

4.2 Standardized National database of Addresses

Let's understand what this means. The same address shows up in multiple ways in different documents e.g.

• Passport Address - 201, Block D, Acropolis, Near Forum Mall, Koramangala, Bangalore-560029

• Driving License Address - D-201, Acropolis, Koramangala, Bangalore-560029

This is even complicated by the fact that most systems do not mandate obtaining Pin code of an address. This is one of the biggest issue world over. In countries like USA, national databases of addresses are maintained which helps in standardization but in a country like India, this seems an impossibility.

4.3 Address Changes

Address, like identity, is a key item that is recorded on most financial and non-financial transactions. Think of any service that you have received recently that did not involve providing your address? Let's understand how this poses a risk of fraud. A customer could report multiple addresses thereby creating multiple identities. This becomes a far complex issue where there is no unique national identity. Even if there was a unique national identity, a new address reported by the same customer on a more recent transaction should also flag the older records for the same customer for a potential update. Failure to do so could increase identity theft risk even if customer didn't intentionally hide the change. It is estimated that about 10% of the population of the world, change their address every year. For a bank like State Bank of India (SBI) which has about 175 million savings accounts [6], it would mean over 50,000 address changes every day!

As it is clear from the above illustration that absence of a national identifier coupled with absence of standardized address database and lack of a method to track and report address changes pose serious risk of identity theft, risk of financial fraud, money laundering and threat to national security as it is known that the terrorist organizations have been known to exploit this loophole.

In India, all the above three are an issue. RBI has issued guidelines to banks to allot Unique Customer Identification Code (UCIC) to all their customers, which will help banks to track transactions of customers. But due to above stated 3 issues, banks are finding it very difficult to relate different accounts/services availed by a customer in a bank and provide a truly unique customer code. The burden of connect/merge all different accounts lies on the customer. A law-abiding customer will do it but not all will choose to do so. This will increase risks for the bank for identity theft and money laundering. Let's now take an example and understand how inadequate, existing KYC systems are, because of the gaps they leave.

As shown in Fig. 1, let’s say a customer opens an account in a bank and provides his PAN card as proof of identity along with passport that provides his proof of residence. Now same client opens another account in the same branch and provides

his passport as proof of identity and his driving license (DL) as proof of residence. Since there is no standard format for capturing the address, it appears differently in both instances. Further the data operator makes a mistake and enters the name as "Anurag" instead of "Anuraj". Since the ID provided is different in both cases, hence for all purposes the same customer can barely be identified. Later(in last column of Fig. 1) this client migrates to Delhi and opens a new account with another branch of the same bank providing Delhi driving license as identity proof. As it is obvious from this example, it is a big challenge for a large bank to connect these various accounts and transactions to one customer.

	New Account	New Account	New Account	New Account	New Account
Fname	Anuraj	Anurag	Reena	Anu	Anuraj
Mname				Raj	
Lname	Soni	Soni	Duggal	Soni	Soni
DOB	18-08-1973	18-08-1973	11-03-1976	08-18-1973	08-18-1973
Id Type	PAN	Passport	PAN	DRL	DRL
Id#	AXIOR2748S	16245B	AQYGH4523D	1837YFG34	TY65290UI67
Address	264, Sigma, Prestige St John Woods Apts, Madiwala, Bangalore - 29	264, Sigma, Prestige St John Woods Apts, Koramangala, Bangalore - 560029	264, Sigma, Prestige St John Woods Apts, Madiwala, Bangalore - 560029	Sigma 264, St John Wood , Koramangala, Bangalore - 560029	18/1 Aksandra, Outer Ring Road, New Delhi-110045

Figure 1: An example showing how the same customer record looks different because of inconsistencies in capturing values

Now if the provided address is same on all the documents, the challenge still lies in

Can the customer be uniquely identified or does each identity proof creates a new unique client, even though the person is the same?

One of the reasons behind this is the use of multiple identities and the fact that all these identities need to be correlated. In addition absence of a standardized address results in even the same address looking very different. Given the size of data it is impossible to clean this before storage.

What if the provided address is different?

Here the customer must be tagged for a detailed KYC review as even if the address has changed, KYC norms require all accounts to show the most recent address.

Similarly one needs to revalidate IDs that expired or may not be valid e.g. driving license in previous state.

What about multiple PAN numbers? Why would someone do that? Yes, to evade taxes.

So why one can't ensure data integrity and flag all non-compliance cases? e.g. in the case above, the moment a different address was provided while opening Account#2, it should have been flagged for review? Why wasn't this done? The answer is simple and as follows:

Any conglomerate having group of companies deals with large number of clients, e.g. State Bank of India has nearly 175 million savings accounts (it opened nearly 29 million in FY13). It suffices the first ***V – Volume*** *of Big Data. It sees about 2,000 transactions a second taking place across its network of close to 14,500 branches - second* ***V – Velocity****. That's nearly 15 billion transactions a year [6]. Large amount of HVT (High Value Transactions) being executed across the country. In addition to this, the information being given as part of KYC is rather unstructured (free-form text in Address fields, scanned copies of identity and residence proofs where data may or may not match with the structured customer database) – third* ***V - Variety****. How does one ensure data integrity over such high volume, velocity & variety? There come Big Data Analytics in the picture.*

5. HOW BIG DATA ANALYTICS SOLVES THESE ISSUES?

In one of the most populous nations in the world, India, it is very challenging to effectively implement KYC norms across all bank branches. Too many identity cards can serve up a crisis. In India, a multitude of options has made establishing one's identity a bit confusing. Does one flash the Aadhaar card, PAN card, driving license, or the passport? Here are the some of the steps (as shown in Fig. 2) to solve these issues and reduce risks in KYC:

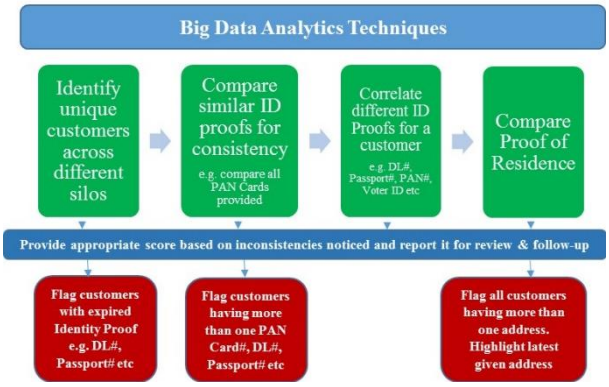

Figure 2: Process Diagram of Risk Identification in KYC in a large Indian Bank

5.1 Identify unique customers

First step is to identify unique customers. For this purpose one could use First Name, Last Name, Date of Birth (DOB), Address and Identification to match unique customers.

5.2 Compare similar id proof for consistency

While opening different accounts if a customer has provided two sets of PAN# or Driving License#, then that customer record should be flagged for further investigation.

5.3 Correlate different ID proofs for a customer

All Identity proofs need to be correlated for a unique customer. It will help to link various accounts opened by a customer and all the subsequent transactions.

5.4 Compare address proof

All proof of residence provided by a customer need to be compared. Assign a score based on inconsistencies noticed. E.g., addresses with different cities or pin code must have a higher score but differences within same pin code could have a lower score. Primary objective here is to determine if the address provided by the customer during various interactions is same/similar or different.

5.5 Report scores above a given threshold for review & follow-up

Operations team could be setup to review records with high score and perform follow-ups with the customer, internal compliance or front-office sales teams. Based on false positives, the system can be tweaked to adjust the scores so that the accuracy could be improved.

Now the outcome of this sample implementation will be shown. In such problems one of the first steps is to generate key-value pairs that are easily comparable. Advantage of using key-value pair is to create strings that are easy to match. **Fuzzy Matching technique** (matching via synonyms, phonetics and approximate spellings) is then applied between data records of customers (Name, ID & Address) across disparate data sets (silos of Banks). Fuzzy matching is an advanced mathematical process that determines the similarities between data sets, information, and facts – where the outcome is neither true nor false, or 100 percent certain, hence the word, "fuzzy." The process compares any data type of any length and from any place in a field to find non-exact matches. For every piece of data examined, the fuzzy matching process will give a probability score to determine the accuracy of the match. For example, 'Rakumar Gupta' might get a 90 percent score of similarity, while 'Raj Gupta' might receive a 75 percent score, as compared to the actual name of 'Rajkumar Gupta'.

The rules shown in Appendix A are used for Name Key Generation to generate key-value pair not only for names but also for addresses. This (modified) technique has been used from best practice guidelines "**Improving the Integrity of Identity Data, Data Matching Better Practice Guidelines"** by Commonwealth Data Matching group [7]. The resulting data set (key-value pairs) from the above example has been shown in Fig. 3.

Fname+DOB	Lname	Id Type	Id#	Address
ANRJ73	SN	PAN	AXIOR2748S	264SGM MDWL BNGLR 29
ANRG73	SN	PASS	16245B	264SGM KRMNGL BNGLR 560029
RYN76	DGL	PAN	AQYGH4523D	264SGM MDWL BNGLR 560029
ANRJ73	SN	DRL	1837YFG34	264SGM KRMNGL BNGLR 560029
ANRJ73	SN	DRL	TY65290UI67	181KSNDR NWDLH 110045

Fig. 3 Resulting Key Value pair after applying the rules mentioned in Appendix A (detailed rules can be found in Best Practice Guidelines in reference [7])

Given the enormity of the data and the speed at which it is coming the authors then used the **MapReduce** technique used by a research paper named "**Social Content Matching in MapReduce**" by Morales, Gionis & Sozio [8]. MapReduce is a programming model for expressing distributed computations on massive amounts of data and an execution framework for large-scale data processing on clusters of commodity servers. It was originally developed by Google and built on well-known principles in parallel and distributed processing dating back several decades. MapReduce has since enjoyed widespread adoption via an open-source implementation called Hadoop, whose development was led by Yahoo (now an Apache project). MapReduce can be used to match customers' data. There can be multiple attributes of a customer identity e.g. Name, Date of Birth, Address, City, Pin Code etc.). They can be assigned different weights (e.g. last name match counts for more than a first name match). For any pair of entities, distance is calculated between corresponding attributes. Attribute wise distances are aggregated over all the attributes of an entity to find the distance between two entities. Using **MapReduce**, data about a customer identity can be exploded and multiple records can be generated with keys from Pin code, city, and last name etc. Grouped on these, and in the reduce phase do a weighted distance calculation to see if any two records were close enough to be considered duplicates. E.g. applying this technique and while checking for similar customers like the one below,

Fname+DOB	Lname	Id Type	Id#	Address
ANRJ73	SN	PAN	AXIOR2748S	264SGM MDWL BNGLR 29

the following scores are received on the data set in question:

Score	Fname+DOB	Lname	Id Type	Id#	Address
70	ANRG73	SN	PASS	16245B	264SGM KRMNGL BNGLR 560029
35	RYN76	DGL	PAN	AQYGH4523D	264SGM MDWL BNGLR 560029
55	ANRJ73	SN	DRL	1837YFG34	264SGM KRMNGL BNGLR 560029
45	ANRJ73	SN	DRL	TY65290UI67	181KSNDR NWDLH 110045

Then a threshold of 50 is chosen for manual review. Though this highlighted two records that belonged to same customer but couldn't catch the last record that belonged to the same customer. But since the customer migrated to another city and used different proof of identification hence detection became difficult unless the threshold was brought down to 45.

The authors also found that for a reduction of every 1 point for threshold, there could be 10x increase in number of records that might need to be manually reviewed in a large data set of over 10 lakh records.

This solution can be further enhanced and will give a credible solution to KYC problem to identify same customers in large Indian banks.

6. REDUCING COSTS AND IMPROVING BANKING EFFICIENCY

While Big Data programs might seem like a hefty investment at the front end, using Big Data analytics internally at banks could drive down the cost of operations by detecting inefficiencies in customer identification, by reducing risks of identity theft and money laundering, by increasing customer satisfaction in a number of different functional operations across the institution.

7. CONCLUSIONS

Hope the readers found this paper informative. There is still lot to learn and try. For those who would like to explore this area a bit more, it would be worth to study in detail various other MapReduce algorithms in associating multiple identities and in finding duplicates. It would also be interesting to see how Hadoop could be used to decentralize processing given that an average financial institution generates tons of data every second. Another area to pursue is the use of Optical Character Recognition (OCR) techniques to accurately decipher scanned documents (hard copies of Identity and Residence proofs) as this is an area that is still processed manually and most often very incorrectly leading to several down-stream challenges. Another completely radical approach to doing KYC is to rely on social media given the amount of information publically available these days. This approach could "build" the KYC information rather than obtain one. One of the inherit advantages of this technique is the absence of bias that is always associated when information needs to be provided by an individual.

8. APPENDIX

A. NAME & ADDRESS KEY GENERATION RULES

1. Remove all non-alphabetic characters from the name. But keep them in the address and put them at the front after removing all characters like "#,&,/"

2. Replace all 'EE' with 'Y'.

3. Replace consecutive repeated characters with one character.

4. Check the first 3 characters and make the following substitutions: 'CHL','CHM','MAC','SCH' is replaced by 'CLL','CMM','MCC','SSS' respectively.

5. Check the last 2 characters and make the following substitutions: 'CE','DT','EE','EH','ET','IE','ND', 'NT','RD','RT','SE','SS','YE','YI','ZE','ZZ' is replaced by ' ','DD','YY','YY','DD','YY','DD', 'DD','DD','DD',' ',' ','YY','YY',' ',' ' respectively

6. Examine the 3 middle characters of the name starting from position 2 and replace them if they match one of the specified substitutions: 'SCH','VSK' is replaced by 'SSS','SSC' respectively .

7. Replace consecutive repeated characters with one character.

8. Drop all vowels.

9. Set the last 2 characters of the NAME-KEY to the year of birth. If unknown set to blank.

9. ACKNOWLEDGMENTS

This is a research paper for 5th National Conference on "Emerging Trends in IT" organized by Christ University, Bangalore. The National Conference would focus on current as well as future challenges and also deliberate on emerging trends in IT.

10. REFERENCES

[1] Talend, "A total data management approach to big data", Research Paper, Oct 2010.

[2] The Data Warehouse Institute. [Online]. Available: http://tdwi.org/portals/big-data-analytics.aspx

[3] Live Mint news (Mar 2013). [Online]. Available: http://www.livemint.com/Politics/hTUpdA8tpufSHI6jfG27gP/Duplicate-Aadhaar-numbers-within-estimates-UIDAI.html

[4] Aadhaar Portal (Jan 2014). [Online]. Available: https://portal.uidai.gov.in/uidwebportal/dashboard.do

[5] Indian government Income Tax site. [Online]. Available: http://www.incometaxindia.gov.in/archive/About%20PAN_06302010.pdf

[6] Business World News Site. [Online]. Available: http://www.businessworld.in/news/finance/banking/the-importance-of-being-sbi/954330/page-1.html

[7] Improving the Integrity of Identity Data - Data Matching Better Practice Guidelines, 2009

[8] Gianmarco De Francisci Morales, Aristides Gionis, Mauro Sozio, "Social content matching in MapReduce" research paper presented at 37th International conference on Very Large Databases (VLDB),2011