Real-time fraud detection: how stream computing can help the retail banking indus...

Business, Industries

Para os meus pais, porque " o valor das coisas nao esta no tempo que elas duram, mas na intensidade com que acontecem. Por isso existem momentos inesqueciveis, coisas inexplicaveis e pessoas incomparaveis" como voces! Obrigado por tudo, Filipe Abstract The Retail Banking Industry has been severely affected by fraud over the past few years. Indeed, despite all the research and systems available, fraudsters have been able to outsmart and deceive the banks and their customers. With this in mind, we intend to introduce a novel and multi-purposetechnologyknown as Stream Computing, as the basis for a Fraud Detection solution.

Indeed, we believe that this architecture will stimulate research, and more importantly organizations, to invest in Analytics and Statistical Fraud-Scoring to be used in conjunction with the already in-place preventive techniques. Therefore, in this research we explore different strategies to build a Streambased Fraud Detection solution, using advanced Data Mining Algorithms and Statistical Analysis, and show how they lead to increased accuracy in the detection of fraud by at least 78% in our reference dataset. We also discuss how a combination of these strategies can be embedded in a Stream-based application to detect fraud in real-time.

From this perspective, our experiments lead to an average processing time of 111, 702ms per transaction, while strategies to further improve the performance are discussed. Keywords: Fraud Detection, Stream Computing, Real-Time Analysis, Fraud, Data Mining, Retail Banking Industry, Data Preprocessing, Data Classi? cation, Behavior-based Models, Supervised Analysis, Semi-supervised Analysis Sammanfattning Privatbankerna har

drabbats hart av bedragerier de senaste aren. Bedragare har lyckats kringga forskning och tillgangliga system och lura bankerna och deras kunder.

Darfor vill vi infora en ny, polyvalent strommande datorteknik (Stream Computing) for att upptacka bedragerier. Vi tror att denna struktur kommer att stimulera forskningen, och framfor allt fa organisationerna att investera i analytisk och statistisk bedragerisparning som kan anvandas tillsammans med be? ntlig forebyggande teknik. Vi undersoker i var forskning olika strategier for att skapa en strommande losning som utnyttjar avancerade algoritmer for datautvinning och statistisk analys for att upptacka bedragerier, och visar att dessa okar traffsakerheten for att upptacka bedragerier med minst 78% i var referensbas.

Vi diskuterar aven hur en kombination av dessa strategier kan baddas in i en strommande applikation for att upptacka bedragerier i realtid. Vara forsok ger en genomsnittlig bearbetningstid pa 111, 702ms per transaktion, samtidigt som olika strategier for att fortsatta forbattra resultaten diskuteras. Acknowledgments " Silent gratitude isn’t much use to anyone" Gladys Bronwyn Stern When I wrote the ? rst words in this report I think I had no idea what a Master Thesis is about!

I can’t blame myself though since I never wrote one before, but if you ask me now to describe this experience I would say that it’s like a road trip: you set yourself a destination, you have a loyal crew that is always there for you, a roadmap, supporters on the side and then the journey begins. Within the latter, you face setbacks with the help of others, you share knowledge, you meet new people and most importantly you get to know them... This journey

would not have been possible without the support, camaraderie and guidance of many friends, colleagues and myfamily.

For all these reasons, I couldn’t let the journey end without expressing my gratitude to each and everyone of them. First and foremost, I would like to express my sincere gratitude to my supervisor, Philippe Spaas, who made it possible for me to work in this project under his supervision and in collaboration with IBM. It was a privilege to work alongside with him and a unique learning opportunity for me! I am indebted for his precious guidance and for the time dedicated not only in helping me understand how a research paper should be formulated, but also in reviewing the latter.

Thank you! I am very thankful as well to Tybra Arthur, who graciously accepted me in her team and supported myinternship, Jean de Canniere who accepted to be my Manager and without whom I wouldn’t have had this opportunity. In this line of thought, I am also grateful to Hans Van Mingroot who helped me secure this project in its negotiation phase. All three were key elements, and their support and guidance throughout the research were important to me and very much appreciated.

I would also like to express my gratitude to Professor Mihhail Matskin at KTH - the Royal Institute of Technology - for having accepted this Master Thesis and for being my examiner. His insights and help were invaluable to achieve more sound end results and put together this ? nal report! In addition, I would like to extend my personal thanks to my Erasmus Coordinator, Anna Hellberg Gustafsson, for her support, kindness and dedication for the duration of the research which was key to the organization of the latter.

She is, for me, the best coordinator I have met and heard about! I would probably not have taken the appropriate steps to have this opportunity within IBM if it weren't for the initial support and guidance of Karl De Backer, Anika Hallier, Anton Wilsens and last but not least Parmjeet Kaur Gurmeet. I truly value their follow-up both on the research and on my experience! On a special note I would like to thank Parmjeet for having been always a good mentor to me and for her support and trust ever since the Extreme Blue internship.

I want to thank each IBMer with whom I came in contact with in the Financial Services Sector Department for welcoming me into their workingenvironmentand for making my stay very enjoyable. In addition to the aforementioned IBMers, among many others and in no speci? c order I would like to thank Daniel Pauwels, Patrick Taymans, Hedwige Meunier, Gauthier de Villenfagne, Michel Van Der Poorten, Kjell Fastre, Annie Magnus, Wouter Denayer, Patrick Antonis, Sara Ramakers, Marc Ledeganck, Joel Van Rossem and Stephane Massonet. It was a real pleasure to share the open space and, more importantly, to meet them!

Dan Gutfreund at IBM Haifa was a key element in the development of this thesis. I am very thankful for the discussions we had about Fraud Detection and for his advice in the different phases that compose this research. In addition, I would like to extend my thanks to Jean-Luc Collet at IBM La Gaude for his valuable help in obtaining a stable virtual machine with InfoSphere Streams. I am thankful to Professor Gianluca Bontempi and Liran Lerman at

Universite Libre de Bruxelles for ? nding the time to discuss about Fraud Detection and Data Mining techniques.

Their insights were vital for the development of the prototype and the overall research. On the same vein, I would like to thank Chris Howard at IBM Dublin for his help in understanding Stream Computing and InfoSphere Streams. His guidance was crucial for a timely comprehension of the ? eld without which I wouldn’t have been able to develop the prototype. I want to thank Mike Koranda and John Thorson at IBM Rochester for their help in understanding the integration of Data Mining and Stream Computing and how to achieve the latter in a more ef? cient manner.

I really appreciated their help with the prototype, especially when atypical errors occurred to more quickly detect the source of the problem. I am also thankful to IBM, as a company, for providing me the opportunity and necessary facilities to conduct my thesis project, as well as to KTH, as university, for having allowed me to take on this experience. I want to take this opportunity to thank my friend, Thomas Heselmans, for having been there ever since the beginning of the research despite my busy agenda. His support and concern were vital in times of greatstressand trouble, thank you for yourfriendship!

The same applies to Stephane Fernandes Medeiros, a great friend of mine who was always there for me and followed my work very closely. In addition, I am thankful to two of my greatest friends, Nicola Martins and Alberto Cecilio, for their friendship, for always supporting me and always having my back. Margarida Cesar is a veryimportant person inmy life, and I would like to

express my gratitude for all the discussions and advice we shared, as well as for the support demonstrated ever since we met. I always take her advice very seriously and she has helped me cope with dif? ulties in more than one occasion, namely during the thesis, and for that I’m very thankful! I am also very grateful to my friend, Arminda Barata, for all the help she provided me in moving and adapting myself to Stockholm. Without her help and concern I wouldn’t have felt at home so easily, and I wouldn’t have liked Stockholm from the very ? rst day. I would like to take advantage of this opportunity to thank all my colleagues and friends in Stockholm for making these two years of study unforgettable, and for shaping the person I am today.

Among so many others, I would like to thank in particular Sanja Jankolovska, Boshko Zerajik, Pedram Mobedi, Adrien Dulac, Filipe Rebello De Andrade, Pavel Podkopajev, Cuneyt Caliskan, Sina Molazem, Arezoo Ghannadian and Hooman Peiro. I couldn’t have made it through without all of them! Last but de? nitely not least, because I didn’t have the chance to formally thank my friends in my previous studies, I would like to take this opportunity to extend my thanks to them for all the good moments we spent together throughout our bachelor degree as well as today.

In particular I would like to thank Miruna Valcu, Rukiye Akgun, Vladimir Svoboda, Antonio Paolillo, Tony Dusenge, Olivier Sputael, Aurelien Gillet, Mathieu Duchene, Bruno Cats, Nicolas Degroot and Juraj Grivna. I reserve a special thank you note to Mathieu Stennier, for both his friendship and support throughout myacademiclife, and for having shared with me what were the best moments I had in Brussels while at University!

I would very much like to express myself in Portuguese to my family so that they can all more easily understand what I have to say, thank you for your understanding: Nao podia deixar de agradecer a toda a minha familia o apoio que demonstraram ao longo deste percurso academico que conhece hoje um novo capitulo. Gostaria de agradecer a todos sem excepcao por acreditarem em mim e nunca duvidarem das minhas capacidades. Obrigado por estarem sempre presentes apesar da distancia, obrigado por se preocuparem comigo e por fazerem com que eu saiba que poderei sempre contar com voces!

Sou verdadeiramente um ser afortunado por poder escrever estas palavras... Um obrigado especial a minha grande avo Olga por estar sempre disposta a sacri? car-se por nos e por telefonar quase diariamente a perguntar se estou bem e se preciso de alguma coisa. Agradeco-lhe do fundo do coracao esse amor que tem pelos netos e que tanta forca transmite! Queria agradecer tambem aos meus primos Rui e Hugo, que sao para mim como os irmaos que eu nunca tive, a forca que me transmitem para seguir em frente face as adversidades da vida. Ambos ensinaram-me imenso durante toda a vida e sao uma fonte de inspiracao constante para mim!

A admiracao que tenho por eles foi como um guia que me levou onde estou hoje... Obrigado por acreditarem em mim para levar a bom porto este projecto e por terem estado sempre presentes a apoiar-me! Gostaria de deixar uma mensagem de apreco ao David, que e mais do que um primo para mim, e um melhor amigo, que sempre esteve presente e sempre se preocupou comigo durante a tese. Foram momentos, frases e situacoes da

vida que ? zeram com que o David se tornasse na pessoa importante que e para mim e ao longo da tese as suas mensagens de apoio foram sempre bem recebidas porque deram-me um alento enorme.

Aproveito tambem para agradecer a minha querida tia Aida e ao meu estimado primo Xico pela preocupacao que tem sempre comigo e por serem uma fonte de inspiracao para mim. Desejo tambem aproveitar esta oportunidade para agradecer a Nandinha e Jorginho todo o apoio que me deram nao so durante estes 6 longos meses mas desde os meus primeiros passos. Sao como uns segundos pais para mim cujo apoio ao longo deste curso e capitulo da minha vida foi primordial. Agradeco, do fundo do coracao, o facto de me tratarem como se fosse um ? lho, por me guiarem e sempre ajudarem! Tenho ainda um lugar especial reservado para o meu tio Antonio.

Um tio que admiro muito, que sempre me quis bem e cujo dom da palavra move montanhas! O seu conselho e para mim uma maisvalia, e agradeco todo o seu apoio e ajuda durante esta investigacao e sobretudo por me guiar quando nao ha estrelas no ceu. Aproveito para vos deixar a todos um pedido de desculpa por nao estar presente como gostaria, e agradeco o facto de que apesar de tudo voces estejam todos de pe ? rme atras de mim! Sem o vosso apoio nunca teria feito metade do que ? z! Costuma-se guardar o melhor para o ? m, e por isso nao podia deixar de agradecer aos meus pais tudo o que ? eram e fazem por mim! A lingua de Camoes e escassa para que eu consiga descrever o quao grato estou... Dedico-vos esta tese, por sempre me terem dado todo o amor, carinho, e ajuda necessaria para ter uma vida

feliz e de sucesso. Deixo aqui um grande e sentido obrigado por terem estado sempre presentes quando mais precisava, por me terem sempre apoiado a alcancar os meus objectivos, por me terem ensinado a viver, a amar, a partilhar e a ser a pessoa que sou hoje. Obrigado! Em particular gostaria de agradecer ao meu pai a compreensao que teve comigo durante este periodo mais ocupado.

Agradecer-lhe a ajuda em conseguir por um meio termo as coisas e a olhar para elas de outro prisma. Agradeco tambem a calma que me transmitiu e transmite, e o apaziguamento que me ensinou a ter face as adversidades da vida. Sem estas licoes de vida, que guardarei sempre comigo, sinto que a tese nao teria sido bem sucedida e eu nunca teria alcancado tudo o que alcancei! A minha mae, agradeco... por onde hei-de comecar? Pela ajuda diaria durante a tese para que os meus esforcos se concentrassem no trabalho? Pela inspiracao diaria de um espirito lutador que nao desmorona face as di? culdades e injusticas da vida?

Agradeco por tudo isto e muito mais pois sem a sua ajuda diaria nao teria conseguido acabar a tese. A admiracao que tenho pela sua forca e coragem ? zeram com que eu tentasse seguir os mesmos passos e levaram-me a alcancar patamares que considerava inalcancaveis! A paciencia que teve durante todo o projecto, mas sobretudo no ? m, e de louvar, e sem o seu ombro amigo teria sido tudo muito mais complicado. Obrigado a todos por tudo! Thank you all for everything! Filipe Miguel Goncalves de Almeida Table of Contents 1 Introduction Part I: Setting the Scene 2 Retail Banking and The State of the Art in Detection and Prevention of Fraud 2. The Retail

Banking Industry . . . . 2. 1. 1 A Short Walk Down Memory Lane . . . . 2. 1. 2 The Retail Banking IT Systems’ Architecture . . . . 2. 2 Fraud . . . . 2. 2. 1 Internet and E-Commerce Fraud . . . . 2. 2. 2 Other Consumer Fraud . . . . 2. 3 Current Solutions . . . . 2. 3. 1 Technology . . . . 2. 3. 2 Analytics and Statistical Fraud-Scoring . . . . 3 Problem De? nition 3. 1 Weak Links in Currently Available Solutions . 3. 1. 1 Bank Card and Pin Code . . . . 3. 1. 2 One-Time-Password or Card Reader . . 3. 1. 3 Biometrics . . . . 3. 1. 4 Analytics and Statistical Fraud-Scoring 3. 2 Facts and Figures . . . . 3. 2. 1 France . . . . 3. 2. 2 United Kingdom . . . . 3. 3 E-Commerce and Internet Banking . . . . 3. 4 Mobile Banking 1 2 3 3 3 4 6 6 12 12 13 14 15 15 16 17 18 18 19 19 19 20 21 22 22 23 23 23 24 24 25 25 28 28

29 . . . . . 30 31 31 31 32 32 33 34 Research Methodology 4. 1 Objective of the Research . . . . . 4. 2 Data Collection . . . . . 4. 2. 1 FICO’s E-Commerce Transactions Dataset . 4. 2. 2 Personal Retail Bank Transactions . . . . . 4. 3 Data Analysis Plan . . . . . 4. 3. 1 Partitioning of the Data . . . . . 4. 4 Instruments and Implementation Strategy . . . . . 4. 4. 1 InfoSphere Streams . . . . . 4. 4. 2 SPSS Modeler . . . . . 4. 4. 3 MySQL Database . . . . . Part II: Behind the Curtains 5 Phase 0: Data Preprocessing 5. Getting to Know the Data . . . . . 5. 1. 1 Attributes and their Types . . . . . 5. 1. 2 Attributes in the Retail Banking Industry and in FICO’s Dataset . . . . . 5. 1. 3 Statistical Description . . . . . 5. 2 Data Reduction . . . . . 5. 2. 1 Dimensionality Reduction . . . . . 5. 2. 2 Supervised Merge and Transformation of Nominal and Categorical Data . 5. 3 5. 4 5. 5 5. 6 . 7 5. 8 Cleaning Process . . . . . . 5. 3. 1 Missing Values . . . . . 5. 3. 2 Noisy Data . . . . . Data Transformation . . . . . 5. 4. 1 Transformation of Times and Dates . . . . . 5. 4. 2 Transformation by Normalization . . . . . Sampling Strategies . . . . . 5. 5. 1 Clustering using K-Means Algorithm . . . . . 5. 5. 2 Under-Sampling Based on Clustering . . . . . Preprocessing Data with Stream Computing . .

. . . . 5. 6. 1 Receiving and Sending Streams of Transactions . . . . 5. 6. 2 Retrieving and Storing Data to a Database . . . . 5. 6. 3 Data Preprocessing using SPSS Solution Publisher . . . . 5. 6. 4 Data Preprocessing using a Non-Generic C++ Primitive Operator Rule-Based Engine . . . . 5. 7. 1 Streams with a Business Rules Management System . . . . Final Thoughts . . . . 36 36 36 37 37 37 39 40 41 42 42 43 45 45 46 48 49 50 51 51 52 53 53 54 55 56 57 57 58 60 60 61 62 62 63 63 66 71 71 73 76 77 6 Phase I: Data Classi? cation 6. 1 Supervised Learning . . . . 6. 1. 1 Ensemble-Based Classi? er . . . . 6. 2 Classi? cation Algorithms . . . . 6. 2. 1 Support Vector Machines . . . . 6. 2. 2 Bayesian Networks . . . . 6. 2. 3 K-Nearest Neighbors (KNN) . . . . 6. 2. 4 C5. 0 Decision Tree . . . . 6. 3 Classi? cation using the Data Mining Toolkit . . . . 6. 3. 1 Weaknesses of the Approach . . . . 6. 4 Classi? cation using SPSS

Modeler Solution Publisher 6. 4. 1 Implementation Details . . . . 6. 5 Model Retraining Architecture: High Level Overview 6. 6 Final Thoughts . . . . 7 Phase II: Anomaly Detection and Stream Analysis 7. 1 Data Aggregation . . . . 7. 2 Bank Customers Aggregation Strategy . . . . 7. 3 Anomaly Detection . . . . 7. 3. 1 Techniques for Anomaly Detection . . . . . . 7. 3. 2 Mahalanobis Distance . . . . 7. 4 Stream Analysis . . . . 7. 4. 1 Window-Based Operators . . . . 7. 4. 2 Window-Based Anomaly Detection Strategy 7. 5 Final Thoughts . . . . Part III: Critical Review 8 Overall Evaluation 8. 1 Performance Measurement Techniques . . . . 8. 1. 1 Performance Metrics . . . . 8. 1. 2 Accuracy Levels . . . . 8. 2 Data Preprocessing and Business Rules Analysis . . . 8. 3 Data Classi? cation . . . . 8. 3. 1 Un-preprocessed Classi? er Analysis . . . . . . 8. . 2 Preprocessed Un-Sampled Classi? er Analysis 8. 3. 3 Preprocessed Sampled Classi? er Analysis . . . 8. 3. 4 Ensemble-Based Classi? er Analysis . . . . 8. 4 Anomaly Detection . . . . 8. 5 Overall Concept . . . . 8. 6 Future Work . . . . 8. 6. 1 Extend Services . . . . 8. 6. 2 eXtreme Scale . . . . 78 78 78 79 80 80 81 83 84 87 88 89 90 91 92 i 8. 7 8. 6. 3 Architecture and Data Mining

Algorithms . . . . Final Thoughts . . . . 93 94 95 i vi 9 Conclusion Appendix A Supporing Figures Glossary List of Figures Figure 1. 1 Figure 2. 1 Figure 2. 2 Figure 2. 3 Figure 2. 4 Figure 2. 5 Figure 2. 6 Figure 2. 7 Figure 2. 8 Figure 2. 9 Figure 3. 1 Figure 3. 2 Figure 3. 3 Figure 3. 4 Figure 3. 5 Figure 3. 6 Figure 3. 7 Figure 3. 8 Figure 4. 1 Figure 4. 2 Figure 4. 3 Figure 4. 4 Figure 4. 5 Figure 4. 6 Figure 4. 7 Figure 5. 1 Figure 5. 2 Figure 5. 3 Figure 5. 4 Figure 5. 5 Figure 5. 6 Figure 5. 7 Figure 5. 8 Figure 5. 9 Figure 5. 10 Figure 5. 11 Figure 5. 12 Figure 5. 13 Figure 5. 14 Figure 5. 15 Figure 6. Figure 6. 2 Figure 6. 3 Figure 6. 4 Figure 6. 5 Lost in Translation . . . . As-Is Banking IT Architecture . . . . Hype Cycle for Application Architecture, 2009 To-Be Banking IT Reference Architecture . . . . MitB Operation . . . . Possible Paypal website (1) . . . . Possible Paypal website (2) . . . . Keyboard State Table method . . . . Windows Keyboard Hook method . . . . Kernel-Based Keyboard Filter Driver method 2 5 5 5 8 10 10 11 11 11 16

16 17 20 20 20 21 21 24 25 26 26 27 28 29 30 32 32 33 34 35 35 36 37 40 40 42 45 46 48 50 51 52 53 54 Components of the Chip and Pin Attack . . . . . Attack to Card Illustrated . . . . . One-Time-Password Hacking Material and Architecture . Number of European Internet Users and Online Purchasers Forecast: US Online Retail Forecast, 2010 to 2015 . . . . . Web Growth has Outpaced Non-Web Growth for Years . . . US Mobile Bankers, 2008-2015 . . . . . US Mobile Banking Adoption . . . . . CRoss-Industry Standard Process for Data Mining . . . . . Streams Programming Model . . . . . " Straight-through" processing of messages with optional storage. Backup and Fail-Over System for Streams . . . . . Multiple-Machines Architecture . . . . . Analytical and Business Intelligent Platforms Compared . . . . . Global Flow of Events: Stream-Based Fraud Detection Solution . Overall SPSS Modeler Stream for the Of? ine Data Preprocessing Phase Frequency of Transactions per Hour . . . . . Amount Transferred per Transaction . . . . . Data Feature Selection in SPSS . . . . . Data Preparation Preprocessing Phase in SPSS . . . . . SPSS Stream CHAID Tree Model . . . . . CHAID Tree for Data Reduction . . . . . Filtering Null Values with SPSS . . . . . Cyclic Values of Attribute hour1 . . . . . K-Means Modeling in SPSS . . . . . Clustering with K-Means in SPSS Modeler . . . . . Stream-based Application: Data Preprocessing and Rule-Based Engine Stream-based Application: Data Preprocessing . . . . . Stream-

based Application: Rule-Based Engine . . . . . Interaction Between a BRMS and a Stream-based Application . . . . . Classi? cation in Stream-Based Application .

Ensemble-Based Classi? er . . . . . Classi? cation in SPSS . . . . . Support Vector Machines (SVMs) Illustrated Example of a Bayesian Network . . . . . i Figure 6. 6 Figure 6. 7 Figure 6. 8 Figure 7. 1 Figure 7. 2 Figure 7. 3 Figure 7. 4 Figure 7. 5 Figure 7. 6 Figure 7. 7 Figure 7. 8 Figure 7. 9 Figure 7. 10 Figure 7. 11 Figure 7. 12 Figure 7. 13 Figure 8. 1 Figure 8. 2 Figure 8. 3 Figure 8. 4 Figure 8. 5 Figure 8. 6 Figure 8. 7 Figure 8. 8 Figure 8. Figure A. 1 Figure A. 2 Figure A. 3 Figure A. 4 Figure A. 5 Figure A. 6 K-Nearest Neighbors Illustrated . . . . . Section of C5. 0 Decision Tree . . . . . SPSS C&DS: Classi? er Retraining . . . . . Anomaly Detection Stream-based Application . . . . . Aggregate Bank Customers . . . . . Learning a classi? er model for the normal class of transactions . . Transaction not belonging to a cluster . . . . . .

Transactions far from the clusters’ center . . . . . Mahalanobis Distance Illustrated . . . . . Mahalanobis Distance: Stream-based Application . . . . . Window Types . . . . . Tumbling Windows . . . . . Sliding Windows . . . . . Partitioned Keyword . . . . . .

. . . . Account average expenses and frequency of transactions in 3 days Window-Based Analysis: Stream-based Application . . . . 54 55 60 61 63 64 65 65 66 67 71 71 72 73 73 74 78 79 84 86 88 89 92 92 94 ii iii iii iv iv v Benchmarking Stream-based Application: Concept for Each Processing Step . . Confusion Matrix . . . . Comparison between Un-Preprocessed and Preprocessed Data: Accuracy Levels Comparison between Sampled Datasets: Accuracy Levels (TP/FP) . . . . Stream Analysis: Debited Account . . . . Overall View of the Solution: Accuracy Levels (TP/FP/FN) . . . . Overall Structure of the Financial Services Toolkit . . . . In-Memory Database with InfoSphere Streams . . . . Stream-Based Application: a Flexible and Multifaceted Architecture . . . . Stream-based Application: Overview . . . . Time per Transaction for each of the Data Preprocessing Approaches . . . . Time per Transaction for Preprocessing the Data and Examine the Business Rules . Metrics Data Classi? cation Process . . . . Anomaly Detection Time per Transaction . . . . Fraud Detection: Time per Transaction . . . . List of Tables Table 3. 1 Table 5. 1 Table 5. 2 Table 6. 1 Table 7. 1 Table 8. 1 Table 8. 2 Table 8. 3 Table 8. 4 Table 8. 5 Table 8. 6 Table 8. 7 Table 8. 8 National fraud in France categorized by transaction type . . . . Communalities PCA/Factor Analysis . . . . Steps for Under-

Sampling Based on Clustering (SBC) . . . . . Supported Mining Algorithms: Data Mining Toolkit . . . . . Hardware Speci? cation . . . . . Individual Classi? er Accuracy Levels - Un-Preprocessed Training Set . . . . . Individual Classi? er Accuracy Levels - Un-Sampled Preprocessed Training Set Multiple Sampling Ratios Analyzed . . . . . Multiple Sampling Ratios Analyzed . . . . . Ensemble-Based Classi? r: Balanced . . . . . Ensemble-Based Classi? er: Maximum Fraud Detection . . . . . Ensemble-Based Classi? er with Mahalanobis: Balanced Model Combination . Ensemble-Based Classi? er with Mahalanobis: Maximizing Fraud Detection . . . . . 19 34 41 56 77 81 83 85 85 87 87 89 89 List of Algorithms Algorithm 1 Algorithm 2 Algorithm 3 Algorithm 4 Algorithm 5 Algorithm 6 Algorithm 7 Algorithm 8 Algorithm 9 Algorithm 10 Algorithm 11 Algorithm 12 Algorithm 13 Algorithm 14 InputSource: Receive Incomming Transactions . . . . . ODBCEnrich: Enrich an Incomming Transaction . . . . . Non-Generic C++ Primitive Operator: Manual Preprocessing . . . . . Preprocessing: Manual Preprocessing of Incoming Transactions . . . . . Functor: Split Stream for Preprocessing and Rule-Based Engine . . . . . Join: Append Business Rules to Preprocessed Transaction . . . . . Join: Append Business Rules to Preprocessed Transaction . . . . . Data Mining Toolkit Operator: Decision Tree C5. 0 Classi? er . . . . . Non-Generic C++ Primitive Operator: Supervised Analysis . . . . . Classi? cation

Ensemble: Constructor() . . . . . Classi? cation Ensemble: process(Tuple & tp, uint32_t port) . . . . . Variance-Covariance Inverse Matrix used in the Mahalanobis Distance . . . Individual Account Anomaly Detection Approach . . . . . Voting Protocol: Mahalanobis Distance, Window-Based and Classi? er Score . . . . . 43 44 45 46 47 47 47 56 58 58 59 68 75 75 Chapter 1 Introduction " A journey of a thousand miles must begin with a single step" Lao Tzu " If you work on fraud detection, you have a job for life". These were the words used by Professor David J.

Hand1 in one of his talks to synthesize the vast research ? eld that is Fraud Detection. Indeed, this ? eld consists of multiple domains, and is continually evolving through time with new strategies and algorithms to counter the constantly changing tactics employed by fraudsters2 . In this line of thought, currently available solutions have been unable to control or mitigate the everincreasing fraud-related losses. Although thorough research has been done, only a small number of studies have led to actual Fraud Detection systems [27], and the focus is typically on novel algorithms aiming at increasing the accuracy levels.

To this end, we want to look at the problem from a different angle, and focus on the foundations for a real-time and multi-purpose solution, based on a technology known as Stream Computing, able to encompass these algorithms while creating the possibilities for further research. We subdivide our study in three main parts. We begin with an overall understanding of the topic being discussed by de? ning the research environment, its problems

and presenting the solutions currently available. In addition, we conclude this ? rst part by both specifying the structure, and outlining the objective of the research.

The second part explores the overall course of action to bring about a Stream-based Fraud Detection solution. From this perspective, we discuss different strategies previously researched in Data Preprocessing, Data Classi? cation and Behavior-based Analysis, and tackle their combination and integration in a Stream-based application. Last but not least, we review the overall solution proposed, and examine the possibilities offered by the latter for further research in the ? eld of Fraud Detection in the Retail Banking Industry. Senior Research Investigator and Emeritus Professor ofMathematicsat the Imperial College of London, and one of the leading researchers in the ? eld of Fraud Detection - http://www3. imperial. ac. uk/people/d. j. hand - link to the presentation: http://videolectures. net/mmdss07_hand_stf/ 2 a person intended to deceive others (i. e. one who commits fraud) [de? ned in the Glossary] 1 Part I: Setting the Scene " Great things are not done by impulse, but by a series of small things brought together" Vincent van Gogh Fraud Detection in itself is interlinked with numerous ? lds of study, and before the play’s main action, we want to set the stage. In order to avoid getting off track and allowing you to better understand the scope, contents, choices made, and requirements of the research, we divided this act in three scenes. In the ? rst, we introduce the main actors - namely banks, bank customers and fraudsters. In addition, we also present the current situation in the Detection and Prevention of Fraud in banks, describing the techniques being used both to counter and to commit

fraudulent transactions. The second scene introduces the overall problem of fraud in the Banking Sector.

It identi? es the weaknesses of the latest solutions, and quanti? es fraud losses as accurately as possible in some European countries and this based on the most recent data. We then take a step further and comment on new trends, and predict possible risks banks might incur from them. Before the end of the act, we introduce the two main parts of the play, as well as how we intend to approach the problem. More precisely, we provide some speci? cs regarding the research conducted, the tools used and the plan followed to reach our conclusions. Figure 1. : Lost in Translation 2 Chapter 2 Retail Banking and The State of the Art in Detection and Prevention of Fraud " There are things known and there are things unknown, and in between are the doors of perception" Aldous Huxley Businessmen and politicians, before sealing deals or taking political decisions, are known to go through a phase of reconnaissance - the military term for exploring enemy or unknown territory. Just as it is important to them, so it is for you when you are about to dive into the speci? cs of a real-time fraud detection solution.

In this line of thought, it is important to grasp the context of the research to better understand the concepts discussed. To do so, we start this chapter with an overall view of the Retail Banking Industry, to understand both its services and IT architecture (Section 2. 1); we continue with a de? nition of fraud together with a description of the different fraud types that affect banks and how they operate (Section 2. 2); lastly, we give an overview of some of the current solutions available (Section 2. 3). 2. 1 The Retail Banking

Industry To describe the banking industry’s evolution that started earlier than 2000 B.

C. [91], deserves almost a research paper on its own. For this reason, and because we don’t want to divert from the topic, we start by solely providing a simple and brief resume about the origins of the banking industry (Section 2. 1. 1). The latter is an interesting talking point that not only allows you to understand how it all started, but also to perceive the challenge of keeping a bank pro? table. Additionally, it is a good introduction to understand a more technical description of the IT architecture behind the banking services (Section 2. 1. 2). 2. 1. 1 A Short Walk Down Memory Lane

It all started with barter back in the time of Dravidian India, passing through Doric Greece to preRoman Italy, when a cow or an ox was the standard medium of exchange. [91] However, given the dif? culty of trading fairly, evaluating different goods with the same standards, and ? nding suitable goods for both parties involved, the invention of " money" inevitably developed. Indeed, the origin of the word money is pecunia in Latin, which comes from pecus, meaning cattle. Through time, money evolved in the different civilizations and became not only a symbol but also a key factor in trading.

Together with the development of the art of casting, the different mediums of exchange evolved gradually from random precious metals to what we now know as currency. This developments made our forefathers the proponents of the ? rst banks for reasons that are still of applicability in today’s banking system. The code of Hammurabi in the early 2000 B. C. stated " If a man

gives to another silver, gold or anything else to safeguard, whatsoever he gives he shall show to witnesses, and he shall arrange the contracts before he makes the deposits. [91] It is therefore clear that the Babylonians already placed back in their time their valuable possessions in a safe place, guarded by a trusted man. 3 Nevertheless, the real inspiration for the banking system as we know it today came from the Greeks. Unlike the Babylonians, the Greeks didn’t have a government and therefore the country was divided into independent states that were constantly either at war or in a state of unrest. [91] In these turbulent times, they found Temples to be the only safe place able to survive the test of wartime.

They were seen as safe deposit vaults, marking the beginning of the functions of our current banks. Indeed, records show that the Temples not only kept money safe but also lent the funds at a certain interest rate. In addition, even though safeguarding the money started as a service free of charge, it soon turned into a business where small commissions were applied. The banking industry continued to evolve through time, from the commercial development of the Jews; passing by the establishment of the Bank of St.

George, the Bank of the Medici and the Bank of England, to the rise of the Rothschilds, and the development of banking in the land of the Vikings. [91] At this moment in time, a major bank is a combination of a dozen of businesses, such as corporate, investment and small business banking, wealth management, capital markets. One among these is the retail banking industry. [46] The retail banking industry is characterized by a particularly

large number of customers and bank accounts in comparison to any other banking business, which results in a much higher number of transactions, services and products.

In addition, it relies more and more on technology due to the levels of cooperation between banks, retailers, businesses, customers leading to an ever-increasing amount of information processing requirements. In a nutshell, today’s banks follow the same principle described earlier by borrowing from clients in surplus and lending to those in de? cit. This triangulation is a win-win situation for the bank and its customers: the bank makes revenue from the net interest income, which is the difference between what it pays to the lending customer and what it receives from the borrower.

Nevertheless, the bank can’t lend all the deposits and needs to guarantee that a certain percentage is kept aside to satisfy customer withdraws and requirements. [92] Even though the situation varies from bank to bank, it is noteworthy to mention that " more than half of a retail bank’s revenue, perhaps three-quarters, comes from this intermediation role in the form of net interest income". [46] To conclude, in today’s world, and after years of evolution, retail banks provide you with a multitude of services for which they charge fees, mainly to cover the maintenance of the infrastructure and the bank’s structure.

These added up together account between 15% to 35% of the net interest income. [46] Among the services you can ? nd payment services, phone banking, money transfer, ATMs1 , online banking, advisory services,

investment and taxation services, mobile banking and many more. How does a bank ef? ciently govern, offer and maintain all these services? 2. 1. 2 The Retail Banking IT Systems' Architecture Just as banking services evolved through time so did the overall back-end architecture allowing a bank to provide all the aforementioned services. This evolution was especially prominent after the unveiling by Barclays Bank f the ? rst ATM machine in 19672 : from that moment on, banks started investing heavily in computerized systems with the goal of automating manual processes in an effort to improve its services, overall status in the market and cut costs. From this perspective, the IT systems of banks matured from the creation of payment systems together with the launch of the international SWIFT network3 in the 70s, to today's core banking system: a general architecture that supports all the channels and services of a bank and where each one of them is digitalized.

An overview of such general architecture is illustrated in Figure 2. 1 [77]. 1 acronym for Automated Teller Machine, a machine that automatically provides cash and performs other banking services on insertion of a special card by the account holder [de? ned in the Glossary] 2 http://www. personal. barclays. co. uk/PFS/A/Content/Files/barclays_events. pdf 3 Society for Worldwide Interbank Financial Telecommunication (SWIFT) is a member-owned cooperative that operates a worldwide standardised ? nancial messaging network through which the ? nancial world conducts its business operations http://www. wift. com 4 This architecture was in place in many banks some years ago, and still is in some cases, but even though it provides the clients with all the necessary banking tools, it had certain drawbacks that

became visible through the modernization and improvement of services. As it is described by both Microsoft [82] and IBM [77]: the as-is architecture has no true enterprise view of a customer because information is duplicated, which leads to inconsistent customer services and promotions across channels; when adding new or changing current products, it takes time to bring Figure 2. : As-Is Banking IT Architecture (source [77]) them to the market and a signi? cant amount of changes to the core system code. This leads to a dif? culty in responding quickly to new challenges and evolving regulatory pressures. Faced with the aforementioned problems, banks had the need to change towards a more ? exible and ef? cient architecture that would allow them to comply with the ever-changing needs of the clients and of the technology. With this n mind, the major players in core banking have switched to a Service-Oriented Architecture (SOA) with the intended goal of improving growth, reducing costs, reducing operational risks, and improving customer experience. [69] [94] [83] [77] [82] As reported by Forrester in a survey in 2007 [82], out of 50 European banks, 53 percent declared they were already replacing their core system while 27 percent were planning to do so and 9 percent had already completed a major transition. The same survey assessed that 56 percent of the banks already used SOA and 31 percent were planning to.

Additionally, in Gartner’s 2009 report (Figure 2. 2 [28]), supports this strategy and believed that SOA-based architectures was increasingly being adopted and would be widely accepted in a time frame of 2 to 5 years. In the latest update (2011th Edition [29]), SOA is entering the Plateau of Productivity, which indiFigure 2. 2: Hype Cycle for Application Architecture,

2009 cates that the mainstream adoption is starting to take off. (source [28]) With this transition to an agile banking platform with a more ? exible product de? ition built on SOA principles, banks expect to gradually simplify their business and become more ef? cient in the long term. Indeed, the aforementioned platform which is illustrated in Figure 2. 3, is meant to provide the banks with faster and easier ways to update the system and comply with changing industry regulations and conditions. Additionally, by having a holistic view of the customer-relevant data across systems, a bank is able to better focus and analyze it with the goal to improve its customers experience by investing in more ef? cient and ? xible customer-centric offerings. Lastly, the architecture allows for integrated customer analytics and insight capabilities. In this line of thought, a stream-based real-time fraud detection solution would be easy to integrate in such an architecture, allowing the bank, as we will see later on, to broaden its services, data analysis capabilities and detect fraud in realtime. Figure 2. 3: To-Be Banking IT Reference Architecture (source [77]) 5 2. 2 Fraud When one wants to get something from others illegally he can do it in two ways: force or trick them into doing so. The ? st is better known as robbery and is usually more violent and noticeable; the second is known as fraud, which is more discrete and therefore preferred by fraudsters. [76] From this we can understand that fraud includes a wide variety of acts characterized by the intent to deceive or to obtain an unearned bene? t. [30] Many audit-related agencies provide distinct insights into the de? nition of fraud that can be brie? y summarized in this way: De? nition 1. Fraud consists of an illegal act (the intentional wrongdoing), the concealment of this act (often only hidden via simple

means), and the deriving of a bene? (converting the gains to cash or other valuable commodity) [30] Given this de? nition, we can further classify the known types of fraud by victim, perpetrator and scheme [76]: • Employee Embezzlement - Employees deceive their employers by taking company assets either directly or indirectly. The ? rst occurs without the participation of a third party and is characterized by an employee who steals company assets directly (e. g. cash, inventory, tools, etc.). In the second, the stolen assets ? ow from the company to the perpetrator through a third party.

Indeed, indirect fraud happens usually when an employee accepts bribes to allow for lower sales or higher purchases prices, or any other dishonest action towards the company. • Vendor Fraud - This type of fraud usually happens when a seller overcharges its products; ships lower quality goods; or doesn’t ship any products to the buyer even though it received the corresponding payment. Vendor fraud happens more frequently with government contracts and usually becomes public when discovered, being one of the most common in the United States. Customer Fraud - Customer fraud takes place when a customer doesn’t pay for the products he purchased, pays too little, gets something for nothing or gets too much for the price. All these situations occur through deception. • Management Fraud - Management fraud, also known as ? nancial statement fraud, is committed by top management who deceptively manipulate ? nancial statements. The interest behind these actions is usually to hide the real economic situation of a company by making it look healthier than it actually is.

However, for the purpose of this research, and given the fact that we are focusing on fraud perpetrated in the retail banking industry, we will mainly focus on every possible bank transaction that a customer can perform. The research will be based in debit, online banking - namely electronic bill payment and giro transfers - and debit plastic card transactions. Fraud that can be perpetrated against these transactions falls within the category known as consumer fraud. Additionally, the latter can be sub-categorized in Internet and e-commerce fraud and other (non-)internet related fraud that we will now describe in more detail. . 2. 1 Internet and E-Commerce Fraud The Internet... a technology that was unknown to many of us 25 years ago and is used now by billions of people either at home, work or on-the-go. We can ? nd webpages from business home pages, to informational wikis, passing throughsocial networkingsites; ? les that take the form of text, audio or video; and a multitude of services and web applications. It took just 3 years for the Internet to reach over 90 million people while the television and the radio took respectively 15 and 35 years to reach 60 million people! 76] This is how fast the medium through which e-commerce fraud takes place has evolved. This informational and technological revolution led to new ways for fraud to be perpetrated while techniques to avoid it have dif? culties to keep up with the pace. Today, businesses depend on the Internet to perform paperless transactions and exchange information between them: they mostly use e-business connections, virtual private networks (VPNs1), and other specialized connections. 76] This type of commerce is known as e-commerce, or electronic commerce, because it takes place over electronic systems. Therefore, even if you think you are not using the Internet, any

operation you make at a local branch, any withdraw you do from an ATM or any purchase you make at a local store with your bank card, a Network transaction takes place. 1 it’s a method employing encryption to provide secure access to a remote computer over the Internet [de? ned in the Glossary] 6

Since most businesses rely on Network-based transactions and, as we will describe later on, Internet users use the network more and more frequently to buy products or services, the North American Securities Administrators Association (NASAA) considers that Internet fraud has become a booming business. [76] With this in mind, there are three standpoints that need to be taken into consideration when describing in more details the risks involved in this category that undermine banks and more importantly their customers: risks lying inside and/or outside the organization.

Risks Inside Banks and Other Organizations The main risks come from within the bank. [76] Indeed, a perpetrator with inside access has knowledge regarding the environment, the security mechanisms and how to bypass them. Additionally, any employee with access to the organization’s network has automatically bypassed ? rewalls and security checks making it easier to in? ltrate systems, steal information or data and cause damage to the bank. From this perspective, the most common example is the superuser access that most IT-related employees (e. g. rogrammers, technical support, network administrators or project managers) have within the company’s infrastructure and database systems. [76] In one survey, " more than a third of network administrators admitted to snooping into human resource

records, layoff lists, and customer databases". [76] A related survey found that " 88 percent of administrators would take sensitive data if they were ? red, and 33 percent said they would take company password lists". [76] Even if a perpetrator does not have personal access to the targeted system and information, there are techniques that he can use to get at them indirectly, i. . via a person of interest: - Snif? ng, also known as Eavesdropping: Snif? ng is the logging, ? ltering, and viewing of information that passes along a network connection. Applications are easily and available for free on the Internet, Wireshark1 and tcpdump2 that allow network administrators to troubleshoot any possible problem in the network. Nevertheless, these applications can as easily be used by hackers to gather information from unencrypted communications. 76] A good example is the usage of unencrypted e-mail access protocols like Post Of? ce Protocol 3 (POP3) or the Internet Message Access Protocol (IMAP) instead of other more secured ones. Since e-mail clients check messages every couple of minutes, hackers have numerous opportunities to intercept personal information. [76] A user could in addition encrypt the body of the email by using Secure/Multipurpose Internet Mail Extensions (S/MIME) or OpenPGP in order to avoid that sensitive information passes through the network in plain text.

Even though security experts have successfully managed to encrypt emails, the reason behind this lack of security is that they have failed to take into consideration the needs of the end-user - namely, " the ability to occasionally encrypt an email without much trouble at all". [113] - Wartrapping: Wartrapping happens when hackers set up free access points to the Internet through their laptops in speci? c locations like airports or inside a company’s

headquarters. Users, unaware that the wi? passes through a hacker’s computer, connect to the latter and navigate the Internet as if they had a secured connection.

When logging their internet banking services and performing transactions, or simply access their emails, the hacker can see the bits and bytes of everycommunicationpassing through any laptop in the clear. In this line of thought, hackers can get caught in their own web as companies are also using what they call honeypot traps. The latter is an information system resource, like a computer, data, or a network site (e. g. wireless entry), whose purpose is not only to divert attackers and hackers away from critical resources, but also to serve as a tool to study their methods. 1] These systems are placed strategically so to look like part of the company’s internal infrastructure even though they are actually isolated and monitored by administrators of the organization. One of the most widely used tools is honeyd3 . [89] 1 2 3 http://www. wireshark. org/ http://www. tcpdump. org/ http://www. honeyd. org/ 7 Passwords are the Achille’s heel of many systems since its creation is left to the end user who keeps them simple and within his or her preferences and life experiences (e. g. birthdays, family names, favorite locations or brands).

In addition, users tend to re-use the same password for different purposes in order to avoid having to remember different ones, which leads perpetrators to gain access to different services and accounts with a single password from the person. In addition, another source of threats are the laptops and mobile devices that many employees take with them outside the company’s

protected environment. While in these unsecured contexts, the devices are exposed to viruses, spyware, and other threats that might compromise again the integrity of other organization’s system once these computers are plugged in the network.

Viruses, trojans and worms are able to enter the protected environment without having to go through ? rewalls and security checks, making it easier to in? ltrate key information systems and bypass defense mechanism. Risks Outside Banks and Other Organizations The Internet not only became a source of services to users and companies but also a rich medium for hackers to gain access to personal systems. Indeed, when performing attacks, hackers are relatively protected because they cross international boundaries - which puts them under a different jurisdiction than the victim of the attack - and are mostly anonymous - making tracking dif? ult. Therefore, the Internet became the defacto technological medium to perform attacks and there are numerous ways of doing so: - Trojan Horses: A trojan horse is a program designed to breach the security of a computer system and that has both a desirable and a hidden, usually malicious, outcome. [86] These programs can be embedded in a bank user’s computer when he views or opens an infected email, visits or downloads a ? le from an unsecured website or even when visiting a legitimate website that has been infected by a trojan. [85] From this perspective, a good example is the man-in-the-browser (MitB) attack, represented in Figure 2. , which uses trojan horses to install extensions or plugins in the browser that are used to deceive a bank customer: Whenever a speci? c webpage is loaded, the Trojan will ? lter it based on a target list (usually online banking pages). The trojan extension

waits until the user logs into his bank and starts to transfer money. When a transaction is performed, the plug-in extracts data from all the ? elds and modi? es the amount and recipient according to the hacker’s preferences through the document object model (DOM1) interface, and resubmits the form to the server.

The latter will not be able to identify whether the values were written by the customer or not and performs the Figure 2. 4: MitB Operation (source2) transaction as requested. [85] - ATM Attack Techniques: An Automated Teller Machine (ATM), is a computerized device that allows customers of a ? nancial institution to perform most banking transactions and check their account status without the help of a clerk. The device identi? es the customers with the help of a plastic bank card, which contains a magnetic stripe with the customer’s information, together with a personal identi? ation number (PIN) code. [2] ATMs are attractive to fraudsters because they are a direct link to customers information and money, and there are security pitfalls with their current architecture [2]: the way data is encoded in the magnetic media makes it easily accessible if a hacker invests some money to buy the easyto-be-found equipment, and time to decode and duplicate the contents; in addition, with a four 1 An interface that lets software programs access and update the content, structure, and style of documents, including webpages [de? ed in the Glossary] 2 www. cronto. com, blog. cronto. com/index. php? title= 2fa_is_dead 8 digit PIN, not only will one in every 10. 000 users have the same number but it also allows brute force attacks to discover the combination. Not to mention the possible physical attacks on ATMs which cannot be considered as fraud (see De? nition 1), there are a couple of ways

fraudsters steal money from bank customers [2]: 1. Skimming Attack: skimming is the most popular approach in ATMs and consists in using devices named skimmers that capture the data from the magnetic strip.

These devices can be plugged in an ATM’s factory-installed card reader and allows for download of all personal information stored on the card. In addition, to obtain the PIN code fraudsters use either shoulder-sur? ng and hidden video cameras, or distraction techniques while the customer uses the ATM. [2] Sometimes fraudsters take a step further and create their own fake teller machines to deceive bank customers; this is considered to be a spoo? ng attack that we will describe in more details below. [39] 2.

Card Trapping: this tech