Datamining para Auditoria de Segurança

Pós-Graduação em Segurança da Informação e Direito no Ciberespaço

V 1.3, V.Lobo, EN 2022

Datamining para Auditoria de Segurança

Victor Lobo

Mestrado em Segurança da Informação e Direito no Ciberespaço

1

Antes da ordem do dia

- Apresentações
  - Victor Lobo, vlobo@novaims.unl.pt
- Trocas de aulas:
  - Dia 14 Março – Aula trocada para 4 de Abril
  - Dia 28 Março – Aula trocada para o fim.
  - Dia 9 Maio - TBD

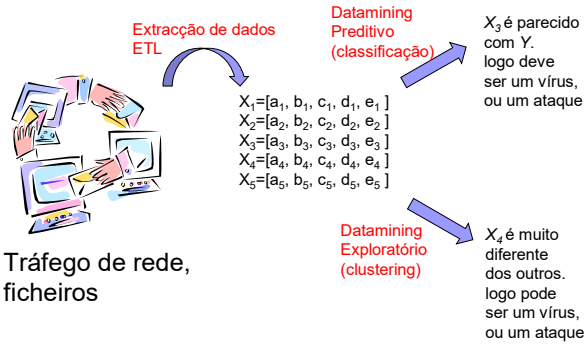
2

Problema:

- Como detectar intrusões, quando não sabemos o que são ? (3º passo no framework NIST)
  - Caso 1: conhecemos casos passados em que foram detectadas intrusões, mas há pequenas variações…
  - Caso 2: conhecemos muitos casos "normais", que variam muito entre si, mas não sabemos o que poderá acontecer de diferente
    - Ficheiros ou ligações normais/anormais
    - Padrões de tráfego normais/anormais

3

Ideia geral



Tráfego de rede, ficheiros

4

Programa (traços gerais)

- Introdução às técnicas para deteção e classificação de cyber-ameaças usando datamining (parte inicial)
- Introdução ao **datamining** e **pré-processamento** de dados
- Técnicas de **visualização** de dados **multi-dimensionais**
- Técnicas de **deteção de outliers** e **comportamentos anormais**
- Técnicas de **classificação** de comportamentos
- Técnicas para deteção e classificação de cyber-ameaças (parte final)

5

Daqui a uns meses, devo…

- Compreender o significado de vectores multidimensionais com dados não necessariamente numéricos, e ser capaz de manipular esses dados
- Compreender como funcionam os principais algoritmos de datamining preditivo.
- Compreender como funcionam os principais algoritmos de datamining exploratório
- Compreender como extrair dados que possam ser usados para detectar anomalias
- Estar feliz por ter aprendido coisas novas !!!

6

Datamining para Auditoria de Segurança

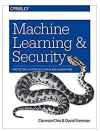
Pós-Graduação em Segurança da Informação e Direito no Ciberespaço

V 1.3, V.Lobo, EN 2022

Bibliografia

- Livros de textos (**não** são seguidos "à risca")
  - Textos de apoio disponíveis no site da UC
  - **Machine Learning and Security: Protecting Systems with Data and Algorithms,** Clarence Chio, David Freeman, O'Reilly Media, 2018

    cap.1,2,3,5
  - **Hands-On Machine Learning for Cybersecurity;** Soma Halder, Sinan Ozdemir, Packt Publishing, 2018



7

Bibliografia

- **Decision Support and Business Intelligence Systems,** Turban, E., J. E. Aronson, et al., Prentice Hall, 2010
- **Data mining: practical machine learning tools and techniques;** Ian H. Witten, Eibe Frank, Mark A. Hall: Morgan Kaufmann, 2011 (WEKA)
- **Python Machine Learning:** Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2**,** Raschka, Packt Pub., 2019

8

Bibliografia mais especializada

- **Data Mining and Machine Learning in Cybersecurity**, Sumeet Dua, Xian Du, ISBN: 978-1439839423, Auerbach Publications, 2011
- **Data Mining Tools for Malware Detection**, Mehedy Masud, Latifur Khan, Bhavani Thuraisingham , ISBN: 978-1439854549, Auerbach Publications 2011.
- **Data Warehousing and Data Mining Techniques for Cyber Security**, Anoop Singhal, ISBN: 978-0387264097, Springer 2006.
- **Applications of Data Mining in Computer Security**, Barbará, Daniel; Jajodia, Sushil (Eds.), ISBN: 978-1-4020-7054-9, Springer 2002.

9

Bibliografia geral de DM

- **Machine Learning**, Tom M.Mitchell, McGraw Hill, 1997
- **Pattern Classification**, Duda, Hart, & Stork, Wiley, 2001
  - **Principles of data mining**, David. J. Hand, Heikki Mannila, Padhric Smyth, MIT Press, 2001
  - **Predictive data mining**, Sholom M. Weiss, Nitin Indurkhya, Morgan Kaufmann, 1997
  - **C4.5:Programs for Machine Learning**, John Ross Quinlan, Morgan Kaufmann, 1992

10

Bibliografia

- Network Traffic Anomaly Detection and Prevention - Concepts, Techniques, and Tools, Bhuyan, Monowar H., Bhattacharyya, Dhruba K., Kalita, Jugal K., ISBN: 978-3-319-65188-0, Springer 2017
- Anomaly Detection Principles and Algorithms, Mehrotra, Kishan G., Mohan, Chilukuri, Huang, Huaming, 978-3-319-67526-8, Springer 2017
- Outlier Analysis, Aggarwal, Charu C., 978-3-319-47578-3, Springer 2017
- Network Intrusion Detection and Prevention - Concepts and Techniques, Ghorbani, Ali A., Lu, Wei, Tavallaee, Mahbod, 978-0-387-88771-5, Springer 2010

11

Resolução de problemas práticos

- MS-Excel
  - Todos conhecem !
  - Resolve a maioria dos problemas simples
- WEKA
  - Java, free, https://www.cs.waikato.ac.nz/ml/weka/
  - Muitos algoritmos, bem documentados
- Orange
  - Python, free, https://orange.biolab.si
  - Interface gráfica
- Outros
  - MATLAB, R, Skikit-learn, Keras.SPSS e Clementine, SAS Enterprise Miner, IBM Intelligent Miner, SAP BI…,

12

Datamining para Auditoria de Segurança

Pós-Graduação em Segurança da Informação e Direito no Ciberespaço

V 1.3, V.Lobo, EN 2022

Repositórios de dados

- **Repositório de Irvine** (UCI)
  - https://archive.ics.uci.edu/ml/index.php
  - Dados, software, artigos
  - Um clássico! Um "must" !
- **Repositório Kaggle**
  - www.kaggle.com/datasets
  - Muito actual, muito activo
- **Repositório do IEEE**
  - IEEE Data Port
  - https://ieee-dataport.org/datasets
- Repositório para Cibersegurança
  - ICSX: http://www.iscx.ca/datasets/ (mas o KDD99 está disponível no UCI)

13

Outros sites interessantes…

- Decisionarium
  - Software GNU, referências, etc
  - http://www.decisionarium.tkk.fi
- DSS Resources
  - Prof. Daniel Power, livros, referências, etc
  - http://dssresources.com/
- Machine Learning Network
  - www.mlnet.org
  - Software, dados, conferências, projectos, etc.
- Fabricantes de soluções "dedicadas"
  - Para gestão de terrenos, para marketing, etc, etc

14

Avaliação (dependente da situação…)

- 1 "Repetição escrita"
  - 40% da nota
- Micro-testes (quizes) e trabalhos de casa
  - 10% da nota
- Apresentação oral e resumo de um artigo
  - 30% da nota
- Projecto de DM para Auditoria de segurança
  - 20% da nota

15

Artigos a apresentar (exemplos… mas **procurem** !)

- Bollmann, C. A., Tummala, M., & McEachen, J. C. (2021). Resilient real-time network anomaly detection using novel non-parametric statistical tests. *Computers & Security, 102, 102146. doi:https://doi.org/10.1016/j.cose.2020.10214*
- Gibert, D., Mateu, C., Planes, J., & Marques-Silva, J. (2021). Auditing static machine learning anti-Malware tools against metamorphic attacks. *Computers & Security, 102, 102159. doi:https://doi.org/10.1016/j.cose.2020.102159*
- Krumay, B., Bernroider, E. W. N., & Walser, R. (2018). *Evaluation of Cybersecurity Management Controls and Metrics of Critical Infrastructures: A Literature Review Considering the NIST Cybersecurity Framework, Cham.*
- Lin, W.-C., Ke, S.-W., & Tsai, C.-F. (2015). CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-Based Systems, 78, 13-21. doi:https://doi.org/10.1016/j.knosys.2015.01.009*

16

Artigos a apresentar (exemplos… mas **procurem** !)

- Mitchell, R., & Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Comput. Surv., 46(4), Article 55. doi:10.1145/2542049*
- Casas, P., Mazel, J., & Owezarski, P. (2012). Unsupervised Network Intrusion Detection Systems: Detecting the Unknown without Knowledge. *Computer Communications, 35(7), 772-783. doi:https://doi.org/10.1016/j.comcom.2012.01.016*
- García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security, 28(1), 18-18-28. doi:10.1016/j.cose.2008.08.003

17

Artigos a apresentar (exemplos…)

- Data Mining for Cyber Security, V.Chandois *et al.*, in Data Warehousing and Data Mining Techniques for Computer Security, Springer, 2006.
- Data mining methods for anomaly detection KDD-2005 workshop report, Margineantu *et al.*, ACM SIGKDD Explorations Newsletter, Volume 7 Issue 2, December 2005.
- On the efficacy of data mining for security applications, Ted E. Senator, ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics -CSI-KDD '09, 2009.
- Metrics for mitigating cybersecurity threats to networks, IEEE Internet Computing, 14, 1, Jan-Fev 2010.
- A Combined Fusion and Data Mining Framework for the Detection of Botnets, Kiayias *et al.*, Conference For Homeland Security, 2009. CATCH '09. Cybersecurity Applications & Technology, March 2009
- A study of Spam Detection Algorithms on Social Media Networks, Jacob Soman Saini, International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013.

18

Artigos a apresentar (…exemplos…)

- Comparative Study of Two- and Multi-Class-Classification-Based Detection of Malicious Executables Using Soft Computing Techniques on Exhaustive Feature Set. Shina Sheen, R. Karthik and R. Anitha; International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013
- Botnets: A Study and Analysis, G. Kirubavathi and R. Anitha, International Conference on Computational Intelligence, Cyber Security, and Computational Models, Coimbatore, India, December 2013
- The VoIP intrusion detection through a LVQ-based neural network, Zheng Lu ; Taoxin Peng, International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Detection of applications within encrypted tunnels using packet size distributions, Mujtaba,G.,Parish, D.J., International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Email classification: Solution with back propagation technique, Ayodele et al. International Conference for Internet Technology and Secured Transactions, 2009. ICITST 2009.
- Malware detection using statistical analysis of byte-level file content, Tabish et al., CSI-KDD '09 Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics, 2009

19

Horário de dúvidas e contactos

- Email: vlobo@novaims.unl.pt
  - Dúvidas e apoio
    - 2ª depois das aulas/5ª Feira PM
    - Por mail em qualquer altura
    - Sempre que estiver disponível …
  - Material de apoio
    - www.novaims.unl.pt/docentes/vlobo
      - Subpasta ***ESCOLA NAVAL,*** subpasta ***Datamining para A.S.***

20

1os trabalhos

- Mini-teste para confirmar conhecimentos e destreza em:
  - Matemática elementar, álgebra, cálculo integral e diferencial, investigação operacional, métodos numéricos, probabilidades e estatística
- 1º Trabalho de casa (para entregar no início da 3ª semana)
  - Propôr um trabalho de projecto de DM-AS. A proposta deverá ter no máximo uma página A4.

21

Conceitos gerais

22

Ideia base:

RECOLHER TODOS OS DADOS POSSÍVEIS !

23

Recolher dados para quê ?

24

Exemplo de Janus

- Olhar o passado e o futuro
- "***Estudar*** o passado para ***compreender*** o presente, e ***prever*** o futuro"

25

Ideias base

Aprender com o passado

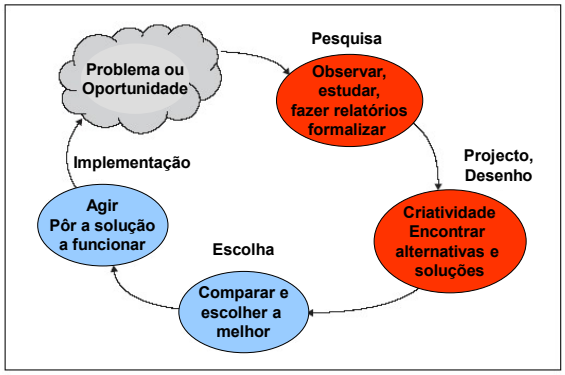
Inferir a partir da experiência

Ferramentas: técnicas de datamining

by any other name…

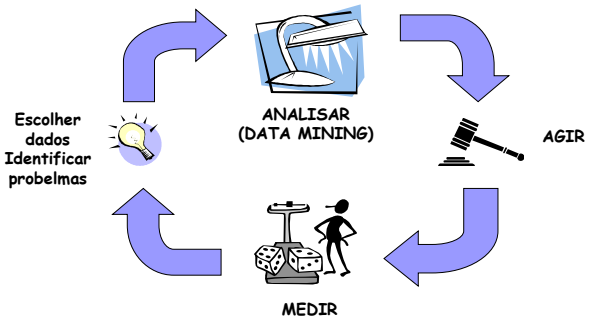
26

Ciclos que nunca terminam…



27

Ciclos que nunca terminam…



28

Simplificando, Datamining é

- A utilização de três técnicas diferentes:
  - Bases de dados
  - Estatística
  - **Aprendizagem máquina**. (Machine Learning)
- Para resolver principalmente dois tipos de problemas
  - Predição
  - Descobrir novo conhecimento

29

Predição e novo conhecimento

- Predição
  - é aprender critérios de decisão para ser capaz de classificar casos desconhecidos
- Descobrir novo conhecimento
  - é encontrar padrões desconhecidos existentes nos dados

30

Tipos de problemas

- **Predição**
  - **Classificação**
  - Regressão
- **Descoberta de conhecimento**
  - **Detecção de desvios**
  - Segmentação de bases de dados
  - **Clustering**
  - Regras de associação
  - Sumarização
  - Visualização
  - Pesquisa em texto

31

Exemplos

- Detecção de fraudes na utilização de um cartão de crédito
- Deferir, ou não, um pedido de crédito
- Prever perdas com seguros
- Prever os níveis de audiência dos canais de televisão
- Classificar os efeitos hidrofónicos produzidos por diferentes navios
- Analisar as respostas de um inquérito médico
- Escolher clientes a quem direccionar uma campanha de marketing
- Cross-selling, fidelização, etc, etc,

32

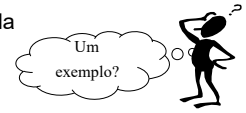
Problemas “a montante”...

- Recolha de dados
- Representação dos dados
- Armazenagem, organização, e disponibilização dos dados
- Pré-processamento dos dados

33

Representação *usual* dos dados

- Representação mais usada = tabela
  - (Existem muitas outras...)
- Exemplo
  - Empresa de seguros de saúde



Variável, característica, ou atributo

Dado, vector, registo ou padrão

Altura	Peso	Sexo	Idade	Ordenado	Usa ginásio	Encargos para seguradora
1.60	79	M	41	3000	S	N
1.72	82	M	32	4000	S	N
1.66	65	F	28	2500	N	N
1.82	87	M	35	2000	N	S
1.71	66	F	42	3500	N	S

34

Modelos versus Dados (ciência versus dataminig?)

- Model based
  - Incorporam o conhecimento à priori
    - F=ma, PV=nRT
    - Conhecimento “certo” pelas “causas”
  - Eventualmente é necessário estimar algum parâmetro (mas poucos)
- Data driven
  - Procuram relações nos dados
    - Relações não implicam causa/efeito
  - Ou não há modelo, ou há um modelo genérico que normalmente é um aproximador universal (com muitos parâmetros)

35

Tipos de dados e operações básicas

36

Dados numéricos

- Inteiros ou reais
- Precisão e gama dinâmica
  - Número de bits
  - Tipo de representação
    - Vírgula fixa, vírgula flutuante, números astronómicos
- Operações
  - Relações de ordem, operações aritméticas
- Exemplos
  - Temperaturas, nº de pessoas, etc
  - 34, 24.5, 20.4x10^{-15}, 32144152353, …
- Dados numéricos multidimensionais
  - Vectores numéricos

37

Dados numéricos

- Como comparar vectores numéricos ?
  - Distâncias d(x,y)
    - 3 condições formais:
      - d(x,y) ≥ 0, $\forall x,y$, *e* d(x,y) = 0, $\Rightarrow x=y$
      - d(x,y) = d(y,x) , $\forall x,y$
      - d(x,y) ≤ d(x,z) + d(z,y) , $\forall x,y,z$
    - Exemplos
      - Distância Euclideana

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$

38

Distâncias entre vectores

- Distâncias de Minkowski de ordem *p*

$$d(x,y) = \left(\sum (x_i - y_i)^p\right)^{1/p}$$

  - *Ordem 1 – Distância de manhatten, ou "city block"*

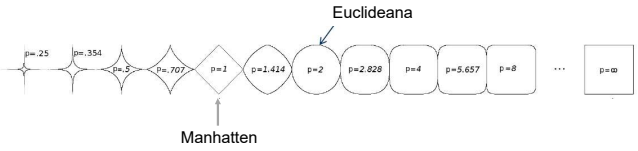
$$d(x,y) = \sum |x_i - y_i|$$

  - *Ordem 2 – Distância Eulideana*
  - *Ordens mais altas*
    - *Dependem cada vez mais da componente mais diferente*
    - *Úteis para evitar "outliers"*

39

Distâncias entre vectores

- Qual a região que está a uma distância de 1 de um dado ponto, usando diferentes índices *p* nas distâncias de Minkowsky num espaço bi-dimensional ?



40

Distâncias entre vectores

- Distâncias ponderadas
  - Dão pesos diferentes a componentes diferentes

$$d(\mathbf{x},\mathbf{y}) = \left(\sum \varphi_i (x_i - y_i)^p\right)^{1/p}$$

  - Se o factor de ponderação fôr a matriz de correlação e a ordem fôr 2, teremos a distância de Mahalanobis, ou distância euclideana normalizada

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{y}-\mathbf{x})}$$

ou simplificando:

$$d(x,y) = \left(\sum_i \frac{|x_i - y_i|^p}{\sigma^2}\right)^{1/p}$$

- Produto interno
  - São uma medida de correlação entre os vectores
  - São a projecção de um vector sobre o outro

$$d(\mathbf{x},\mathbf{y}) = \sum x_i y_i$$

41

Distâncias entre vectores

- Máxima correlação

$$d(x,y) = \max_k \sum x_i y_{i-k}$$

- Cosenos directores
  - É sensível à relações entre as componentes e não à sua magnitude

$$d(x,y) = \cos\theta = \frac{\sum x_i y_i}{\|x\| \times \|y\|}$$

- Outras
  - Menor diferença
  - Maior diferença
  - Tanimoto (aplicado a reais)

$$d(x,y) = \frac{\sum x_i y_i}{\|x\|^2 + \|y\|^2 - \sum x_i y_i}$$

42

Dados categóricos

- Booleanos
  - Só têm valor 0 ou 1
  - Exemplos
    - Tem a altura mínima, tem um curso, tem...
- Ordinais
  - Têm um número finito de valores
  - Os valores têm uma relação de ordem (mas não podem ser feitas operações aritméticas)
  - Exemplos
    - Escalões de vencimentos, Escalas de comportamento
    - Mau/Suficiente/Bom/Muito Bom, Alto/médio/baixo…
- Categóricos (puros)
  - Não têm relação de ordem
  - Exemplos
    - Naipes de cartas, raças,
    - Paus/Ouros/Espadas/Copas, Marinha/Administração Naval/Fuzileiros/…

43

Distâncias entre vectores categóricos

- Distância de Hamming
  - Número de bits diferentes
  - Equivalente à distância de manhatten ou ao quadrado da distância euclideana
  - Exemplo
    - D(0010, 1010)=1, D(0010,1101)=4
- Distância de edição ou de Levenshtein
  - Número de alterações (apagar um valor ou acrescentar um valor)
  - Exemplo
    - D(ABC,AB)=1, D(ABC,AD)=3

44

Distâncias entre vectores categóricos

- Tabela de contingência entre valores dos vectores

		Object x		
		1	0	sum
Object y	1	***a***	***b***	a+b
	0	***c***	***d***	c+d
	sum	a+c	b+d	a+b+c+d

- Métricas:

Coefficients	Equation	Range
Simple Matching (Sokal and Michener 1958)	$\frac{a+d}{a+b+c+d}$	[0,1]
Russel and Rao (Russel and Rao 1940)	$\frac{a}{a+b+c+d}$	[0,1]
Rogers and Tanimoto (Rogers and Tanimoto 1960)	$\frac{a+d}{a+d+2(b+c)}$	[0,1]
Hamann (Hamann 1961)	$\frac{(a+d)-(b+c)}{a+b+c+d}$	[-1,1]
Ochiai II (Ochiai 1957)	$\frac{ad}{\sqrt{(a+d)(a+c)(d+b)(d+c)}}$	[0,1]
Sokal and Sneath (Sokal and Sneath 1963)	$\frac{2(a+d)}{2(a+d)+b+c}$	[0,1]

Coefficients	Equation	Range
Jaccard (Jaccard 1901)	$\frac{a}{a+b+c}$	[0,1]
Anderberg (Anderberg 1973)	$\frac{a}{a+2(b+c)}$	[0,1]
Czekanowsky / Sorensen-Dice (Dice 1945)	$\frac{2a}{2a+b+c}$	[0,1]
Kulczynski I (Kulczynski 1927)	$\frac{a}{b+c}$	[0,+∞]
Kulczynski II (Kulczynski 1927)	$\frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right)$	[0,1]
Ochiai (Ochiai 1957)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0,1]

45

Medidas de semelhança/dissemelhança

- Não obedecem às 3 condições das distâncias
  - Podem não ser simétricas
  - Podem ser o inverso de uma distância
  - Podem não respeitar a desigualdade triangular
- Exemplos
  - Algumas das métricas do acetato anterior
  - "Distância" de Kullback–Leibler

$$d(x,y)=\sum x_i \log \frac{x_i}{y_i}$$

46

Outros tipos de dados

- Conjuntos
  - Podem ser semelhantes a dados categóricos
    - Representados e manipulados como categóricos
  - Podem ser conjuntos de pontos
    - Representados como listas
    - Distância de Hausdorff
      - Maior das menores distâncias de um conjunto ao outro

$$d(x,y)=\max(\min_j d(x_i,y_j))$$

- Árvores ou outros grafos
- Mapas
- Etc,etc,etc…

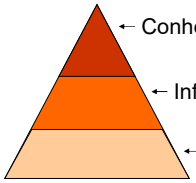
47

Organização dos dados

48

Informação é poder...

- "Água é vida"...
  - Todos os anos morre gente afogada...
- É necessário "trabalhar" a informação
- Hierarquia de compreensão e utilidade



Compreensão "Visual Analytics"

Fusão de dados
Meta-dados
Modelos

Dados em burto.
Aquisição de dados
Redes de sensores

49

SI Operacional *vs* Analítico

- Sistema de Informação **Operacional**
  - Ligado directamente aos **processos**
  - Processamento em tempo real, **contínuo**
  - Muitos dados, **pouco processamento**
  - Constante **mutação**
  - Dia a dia da operação
- Sistema de Informação **Analítico**
  - Ligado aos **decisores**
  - Processamento "off-line", em **tempo diferido**
  - Muitos dados e **MUITO processamento**
  - Maior **estabilidade**
  - Memória da organização

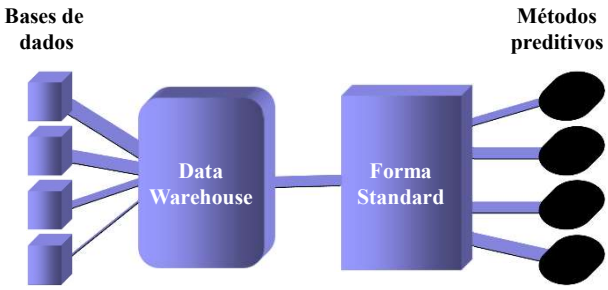
50

Datawarehouse

- Definição de W.H.Inmon
  - *A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.*

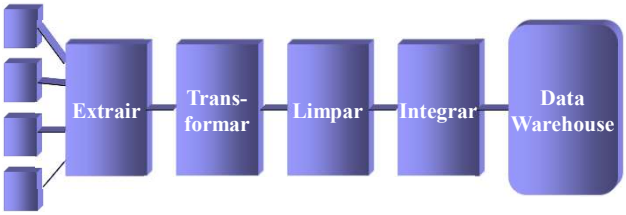
51

O modelo de "data warehouse"



52

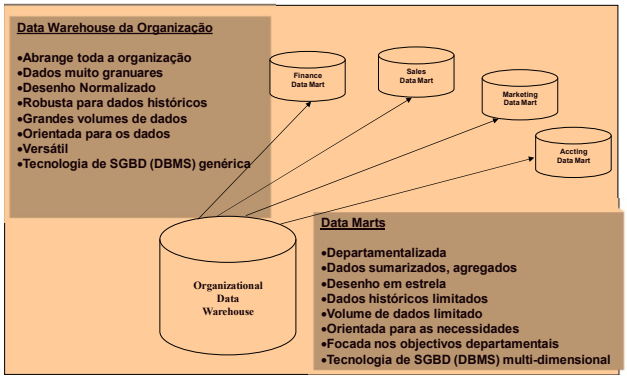
Passos para construir a "data warehouse" (processo de ETL)



Bases de dados

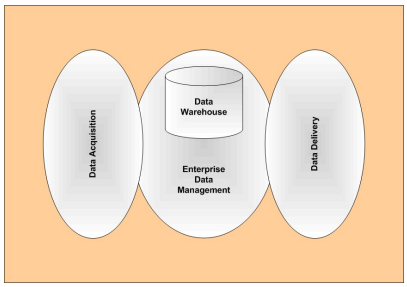
53

Datawarehouse & data-marts



54

Outras perspectivas….



55

Medição, indicadores, visualização

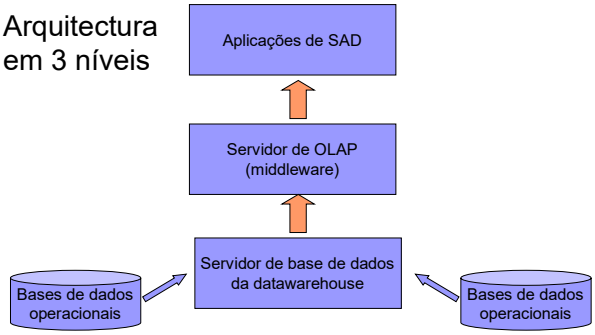
- Relatórios “tradicionais”
  - Relatórios contabilísticos, tabelas de resultados
- **Dashboards**
  - Conceito de “tableau de bord”
  - Um (ou mais) números que indicam a “saúde” da empresa

- **Scorecards**
  - Metodologias para medir “o que é importante” num dado negócio
  - Técnicas para elaboração de “***balanced scorecards***”

56

Acesso à datawarehouse

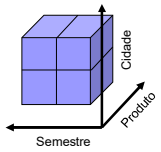
- Arquitectura em 3 níveis



57

Sistemas de OLAP

- OLAP- On-Line Analytical Processing
  - Disponível para muitos sistemas de bases de dados
  - Conjunto de ferramentas de “reporting”: fáceis e flexíveis
- Conceito de **hipercubo de dados**
  - Agrupar segundo **diversas dimensões**
    - Tempo, Local, Produto, Cliente, etc.
  - **Cortes** (slices) e **vistas**
    - Ver o hipercubo sob uma dada perspectiva
    - “Colapsar” (ou não) algumas dimensões
  - **Roll-up**:
    - Consolidar ou agregar em dados mais gerais
  - **Drill-down**:
    - Separar em nódulos mais específicos
  - Outras:
    - Ranking, Filtering, Dicing, estruturas ROLAP, HOLAP



58

Exemplo de um cubo de dados

- dados de vendas por semestre, por produto e por cidade:

Semestre	*Vendas*
Primeiro	16.000,00
Segundo	16.000,00

Produto	*Vendas*
Banana	16.000,00
Laranja	16.000,00

Cidade	*Vendas*
Lisboa	16.000,00
Porto	16.000,00

59

Exemplo de um cubo de dados

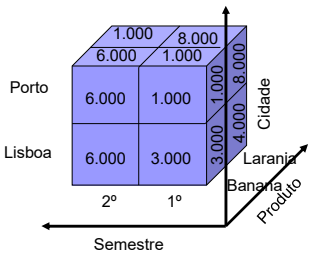
- Dados mais detalhados: numa tabela

Semestre	*Produto*	*Cidade*	*Valor*
Primeiro	Banana	Lisboa	3.000,00
Primeiro	Banana	Porto	1.000,00
Primeiro	Laranja	Lisboa	4.000,00
Primeiro	Laranja	Porto	8.000,00
Segundo	Banana	Lisboa	6.000,00
Segundo	Banana	Porto	6.000,00
Segundo	Laranja	Lisboa	3.000,00
Segundo	Laranja	Porto	1.000,00

60

Exemplo de um cubo de dados

- Dados mais detalhados: num cubo



61

Bibliografia

- George Marakas, Modern Data Warehousing, Mining, and Visualization, Prentice-Hall 2003
- Barry Devlin, Data Warehouse – from Architecture to Implementation, Addison-Wesley, 1997

62

Bibliografia (artigos)

- Kouzes et al., The changing paradigm of Data-Intensive computing, IEEE Computer, Jan 2009

63