Business Intelligence
Cursul 5

Prof. Bologa Ana-Ramona
ASE, Bucuresti

Agenda

1. **Metadate**
2. **Modelarea datelor in BI**
   - QPM- Qlikview Project Methodology
3. **Advanced analytics (in memory, big data)**

1. Metadate

- Date care descriu datele (ex. Dictionarul de date)
  - **Metadate tehnice**: definesc tipurile si structura datelor
  - **Metadate de business**: ofera informatii despre sursa, acuratetea si fiabilitatea datelor
- Exista **6 tipuri de depozite de metadate**
  1. **Schema bazei de date**
  2. **Specificatii ale interfetelor obiectelor (IIOP)**
  3. **Specificatii de transformare a mesajelor** (XSLT)
  4. **Metadatele depozitului de date** (importante pentru analiza datelor)
  5. **Metadate despre cunostinte**
  6. **Specificatiile schemei XML** (eventual pe Internet)

Metadatele depozitului de date

- In DW:
  - pt sursele de date,
  - pt programele si regulile ETL,
  - pt structura datelor
  - pt continutul propriu-zis al DW
- Importanta
  - stabilesc contextul DW (localizarea si intelegerea datelor)
  - usureaza procesul de analiza (identificarea, obtinerea, interpretarea si analiza datelor)
  - sunt o forma de auditare a transformarii datelor (incredere daca se cunoaste CUM au fost obtinute)
  - mentin si cresc calitatea datelor (definire de valori valide)

Tipuri de metadate (dupa destinatie)

- ***Metadate administrative*** - descrieri ale:
  - BD sursă şi ale continutului,
  - obiectelor DW
  - regulilor pentru a transforma datele din sistemul sursă în depozit
  - programe şi instrumente back-end,
  - reguli şi formule de calcul,
  - reguli de securitate şi de acces
- ***Metadate pentru utilizatorii finali*** -rolul de a ajuta utilizatorii să-şi creeze **propriile lor interogări** şi să **interpreteze rezultatele**
  - definitiile datelor din depozit, descrierea lor,
  - rapoarte şi interogări predefinite,
  - definitiile ierarhiilor,
  - calitatea datelor, istoricul încărcării depozitului de date, reguli de eliminare
- ***Metadate pentru optimizare -*** rolul de a creşte performantele depozitului de date.
  - Ex: definitiile agregărilor şi colectii de statistici.

Metadate administrative: campuri

Exemplu de metadate pentru câmpurile depozitului de date

- **Denumire câmp.** Denumirea câmpului, aşa cum va fi folosită în tabela fizică.
- **Titlu.** Denumirea câmpului, aşa cum va apare pentru utilizatori.
- **Tipul datei.** Tipul de date pentru câmpul respectiv, suportat de sistemul de gestiune al bazelor de date.
- **Tipul indexului.** Tipul de index care va fi folosit pentru acest câmp.
- **Key?** Specifică dacă este un câmp cheie în tabela dimensiune.
- **Format.** Formatul câmpului.
- **Descriere.** O descriere sau o definiţie a câmpului.

Metadate administrative: dimensiuni

Exemplu de metadate pentru tabelele dimensiuni ale unui depozit de date

- **Denumire.** Numele tabelei fizice care va fi folosită în depozitul de date.
- **Titlu.** Numele dimensiunii; acest titlu va fi folosit de utilizatori pentru a face referire la dimensiune.
- **Descriere.** Definiția standard a dimensiunii.
- **Tipul folosirii.** Indică dacă un obiect al depozitului de date este folosit ca faptă sau ca dimensiune.
- **Sursa.** Indică sursa primară de date pentru dimensiune.
- **Online?** Indică dacă tabela fizică este populată corect și dacă este disponibilă pentru utilizatorii depozitului de date.

2. Modelarea datelor

Modelul Logic. Modelul Fizic

Modelarea datelor

- Modelul: reprezinta vizual **natura datelor**, **regulile de business** respectate de date si **cum vor fi utilizate** in baza de date:
  - *Modelul conceptual – reprezinta entitatile de business*
  - *Modelul logic- reprezinta logic cum sunt conectate entitatile*
  - *Modelul fizic– realizarea tehnologica ale primelor doua modele*
- are doua parti esentiale:
  1. **Proiectare logica**
  2. **Proiectare fizica**
- modelul datelor nu va include toate datele si codul din baza de date, dar va avea obiecte de tip:
  - tabela,
  - coloana,
  - restrictie,
  - relatie

Ciclul de modelare a datelor

1. ***Colectarea cerintelor de business***
   - interactiune cu analistul de business si utilizatorii finali pt *cerintele* ***de raportare***
2. ***Modelarea conceptuala a datelor***
   - identificarea entitatilor principale si a relatiilor dintre ele
3. ***Modelarea logica a datelor***
   - reprezinta toate cerintele de business , extinzand modelul conceptual cu atribute, chei, relatii, text descriptiv
4. ***Modelarea fizica a datelor***
   - model complet ce include tabele, coloane, relatii, proprietati fizice
5. **Crearea bazei de date**
   - entitati->tabele, atribute -> coloane, tipuri de date, restrictii , indecsi

Pasi pentru crearea **modelului logic**

1. **Identificarea** cerintelor de business
2. **Analiza** cerintelor de business
3. Crearea **modelului conceptual** al datelor. Aprobarea lui de catre reprezentantii de business
4. Crearea noului **model logic de date** care include urmatoarele:
   - Selectarea BD tinta (pt generare scripturi pentru schema fizica)
   - Crearea unui document **cu abrevieri standard** pentru obiectele logice/fizice
   - Crearea **domeniilor**
   - Crearea **regulilor (restrictiilor)**
   - Crearea **valorilor implicite**
   - Crearea **entitatilor** si adaugarea de definitii
   - **Asignarea tipurilor** de date/domeniilor pt atribute
   - Adaugarea de **restrictii CHECK**/reguli sau **valori implicite**
   - Crearea de **chei primare** sau **unice**
   - Crearea **indecsilor**
   - Daca e necesara, crearea subtipurilor si supertipurilor (**mostenire**)
   - Identificarea **relatiilor** intre entitati si crearea **cheilor externe**
   - **Validarea** modelului de date
   - **Aprobarea** modelului logic

Pasi pentru crearea **modelului fizic**

1. Crearea **modelului fizic** pe baza modelului logic
2. **Adaugarea de proprietati** specifice bazei de date in care se realizeaza stocarea (organizare, indecsi, stocare, securitate)
3. **Generarea scripturilor SQL** din modelul fizic; trimiterea lor catre DBA
4. **Compararea** bazei de date cu modelul datelor
5. Crearea unui **document de log** pentru urmarirea modificarilor modelului

- In **transformarea** model logic ->model fizic, **tipurile de date** pot fi complet diferite, conform cerintelor de raportare si restrictiilor fizice (lungimea numelor tabelelor, numelor coloanelor etc)
- **STANDARDIZARE** in modelul logic datelor (aceeasi denumire, tip, abrevieri)

Figure 1. A simple logical data model.

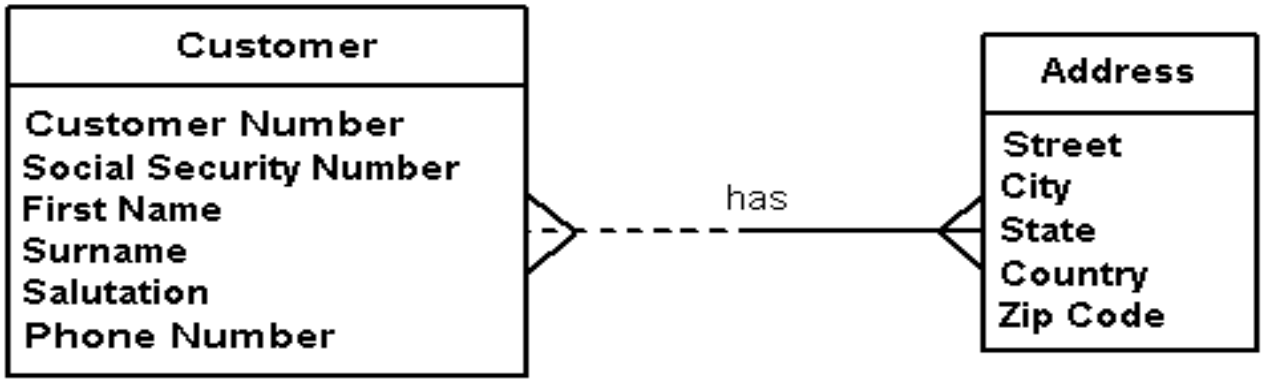
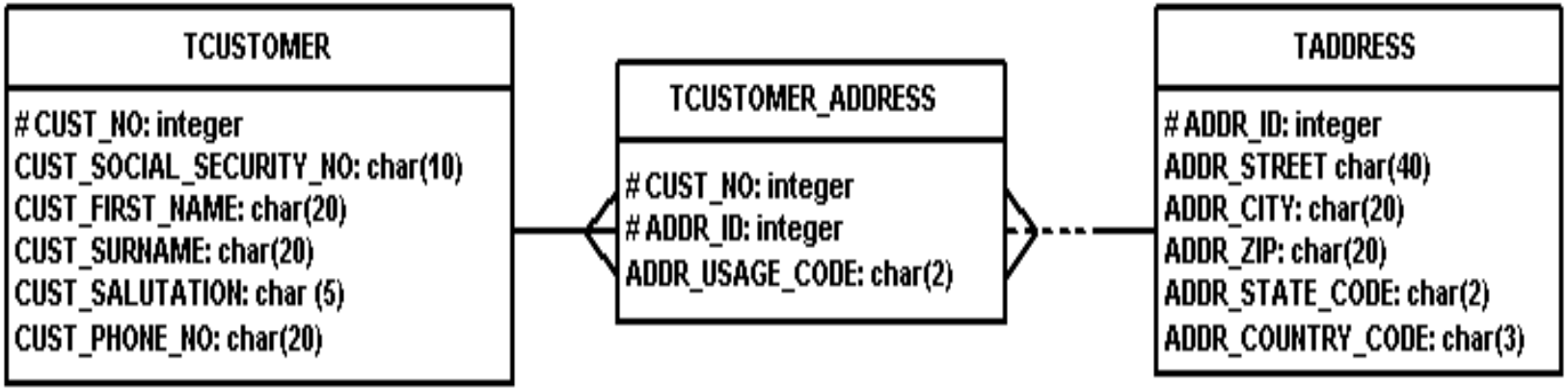
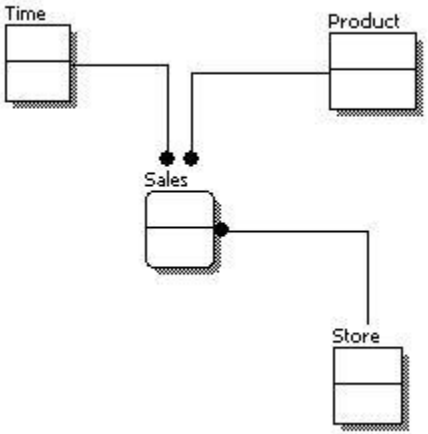
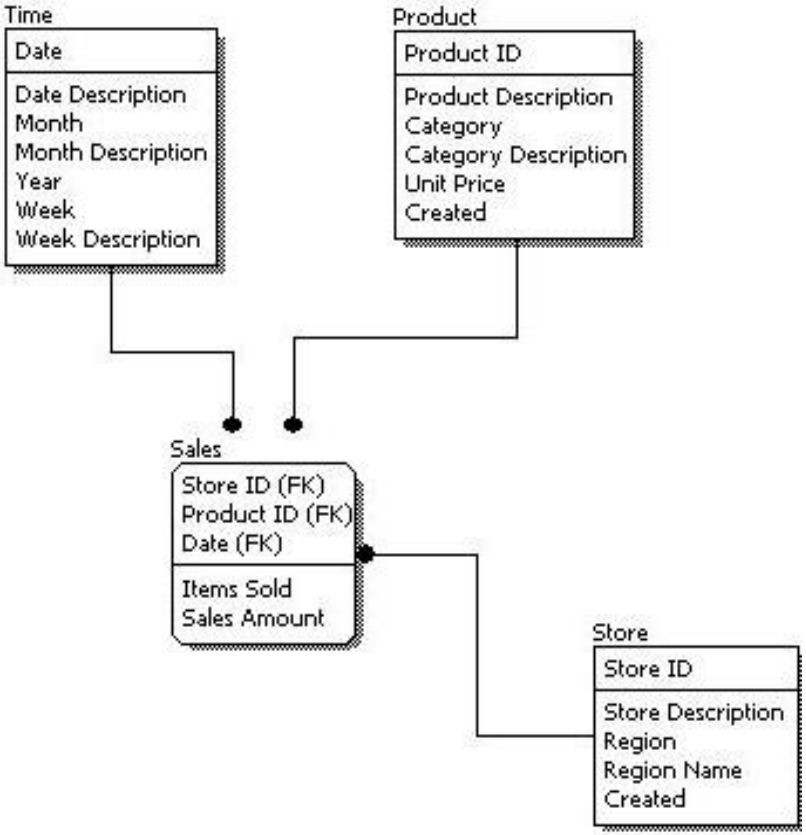
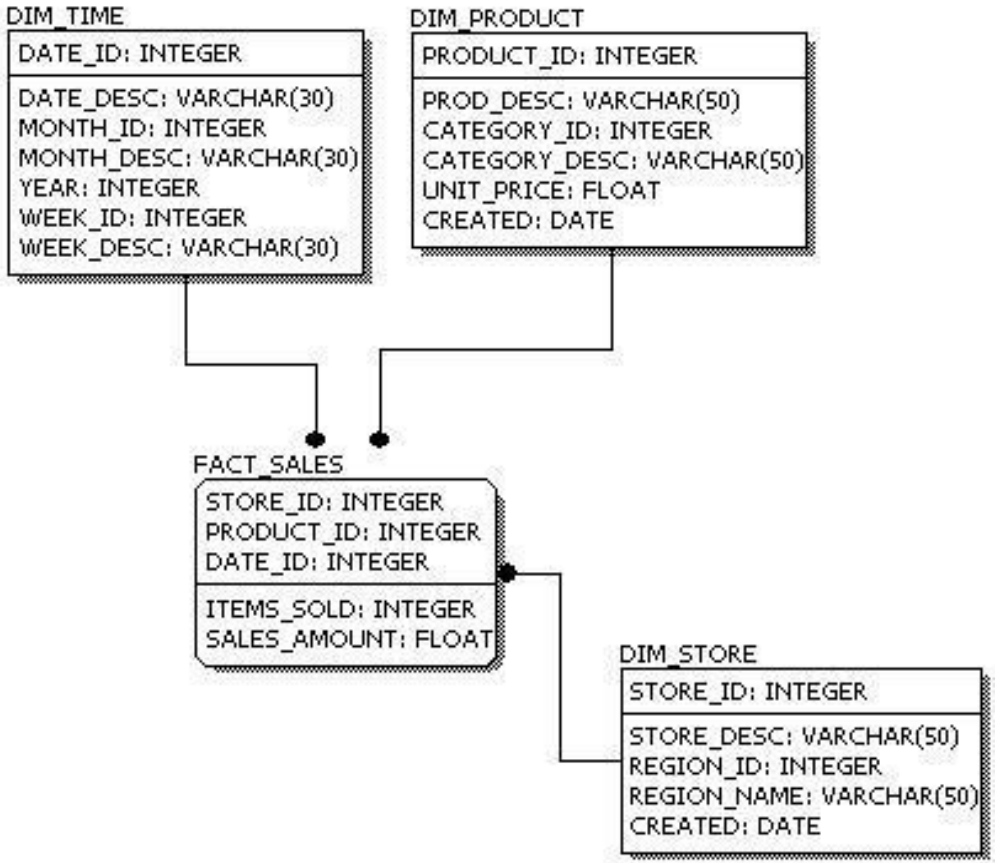


Figure 2. A simple physical data model.



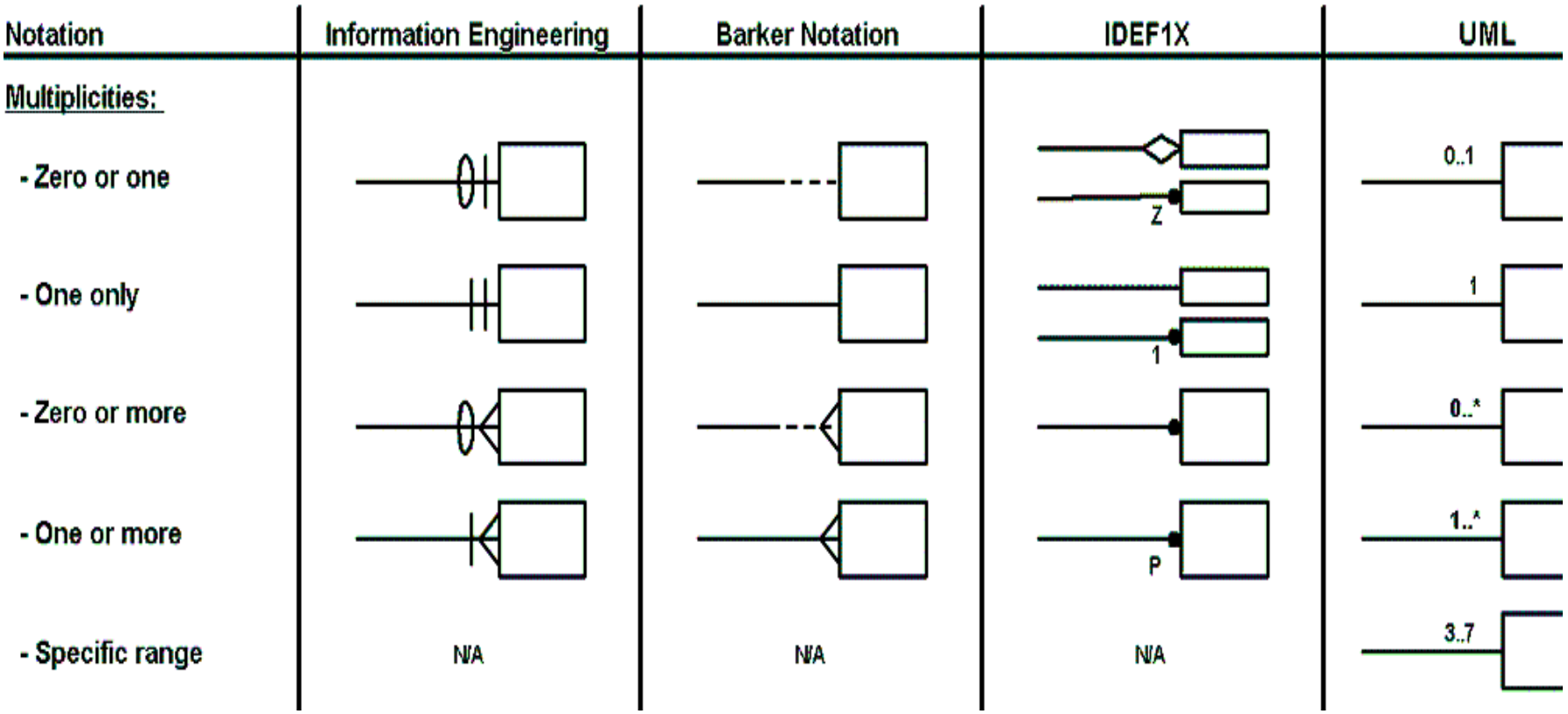

Time
Product
Sales
Store


Time
Date
Date Description
Month
Month Description
Year
Week
Week Description
Product
Product ID
Product Description
Category
Category Description
Unit Price
Created
Sales
Store ID (FK)
Product ID (FK)
Date (FK)
Items Sold
Sales Amount
Store
Store ID
Store Description
Region
Region Name
Created


DIM_TIME
DATE_ID: INTEGER
DATE_DESC: VARCHAR(30)
MONTH_ID: INTEGER
MONTH_DESC: VARCHAR(30)
YEAR: INTEGER
WEEK_ID: INTEGER
WEEK_DESC: VARCHAR(30)
DIM_PRODUCT
PRODUCT_ID: INTEGER
PROD_DESC: VARCHAR(50)
CATEGORY_ID: INTEGER
CATEGORY_DESC: VARCHAR(50)
UNIT_PRICE: FLOAT
CREATED: DATE
FACT_SALES
STORE_ID: INTEGER
PRODUCT_ID: INTEGER
DATE_ID: INTEGER
ITEMS_SOLD: INTEGER
SALES_AMOUNT: FLOAT
DIM_STORE
STORE_ID: INTEGER
STORE_DESC: VARCHAR(50)
REGION_ID: INTEGER
REGION_NAME: VARCHAR(50)
CREATED: DATE

Notatii

- **Notatii pentru modelarea datelor:** Information Engineering (IE), Barker, IDEF1X, Unified Modeling Language (UML)



Implementare Data Warehouse

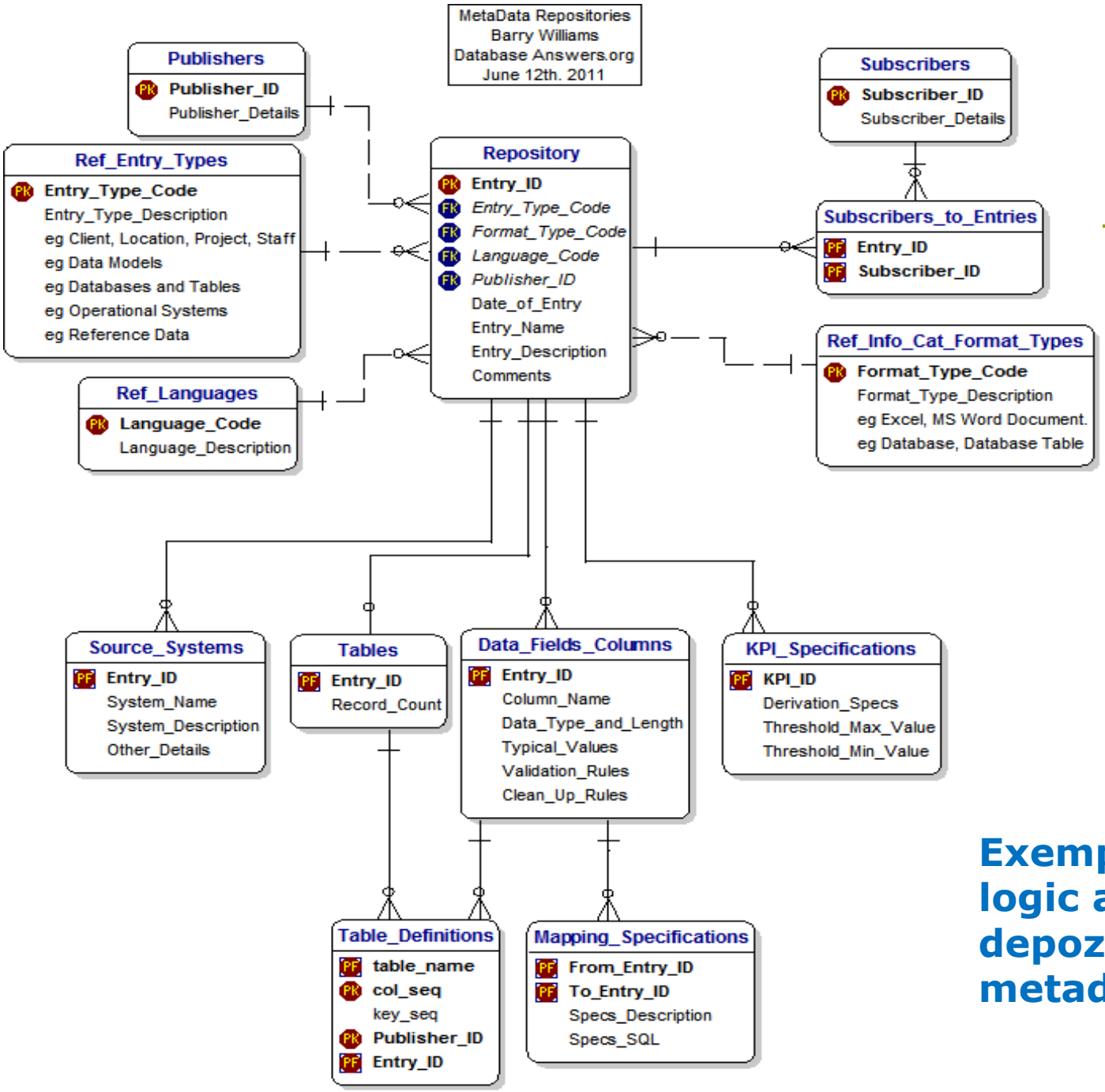
- **Strategii de implementare**
  - Strategie de **tip organizatie** / top – down / metodologie Inmon
  - Strategie de **tip Data Mart** / bottom – up / metodologie Kimball
  - Aplicate corect, ambele strategii conduc la o implementare corecta de Data Warehouse

Depozit de modele de date

- **Modelele datelor** si **metadatele** referitoare la acestea sunt stocate intr-un **Data Model Repository** – acces concurent, pe baza de privilegii
- **Business metadata**
  - **text aditional** , definitie a unui termen (tabela, coloana) asigura intelegerea comuna a semnificatiei
  - util in generarea rapoartelor atat pentru echipa tehnica, cat si pentru non-tehnica,
  - **Metadate TABELA** – numele sistemului sursa, numele entitatii sursa, regulile de business pentru transformarea tabelei sursa, utilizarea tabelei in rapoarte
  - **Metadate COLOANA** – coloana sursa, regulile de business pentru transformarea coloanei sursa, utilizarea coloanei in rapoarte

Exemplu de Business Metadata

Entity (Table) Name	Attribute (Column)	Attribute Definition
TARGET AUTO LOAN BY WEB	Auto Loan Identifier	The number that uniquely identifies an AUTO LOAN.
	Auto Loan Amount	The amount of auto loan that has been approved. Mapping: SOURCE_AUTO_LOAN_BY_WEB.AUTO_LOAN_AMOUNT
	Auto Loan Broker Commission Amount	The commission amount that has to be paid to AUTO loan broker. Note: This column is a derived column and not found in the source system. Derivation Rule: Auto Loan Amount * .01
	Auto VIN Identifier	This column identifies the Auto VIN Number
	Borrower Full Name	The full name of the borrower. Note: This column is a derived column and not found in the source system. Derivation Rule: SOURCE_AUTO_LOAN_BY_WEB.(BOR_FST_NAME concatenated with BOR_LAST_NAME)
	DateTimeStamp	The date on which the record has been created or updated.



Exemplu de model logic al unui depozit de metadate

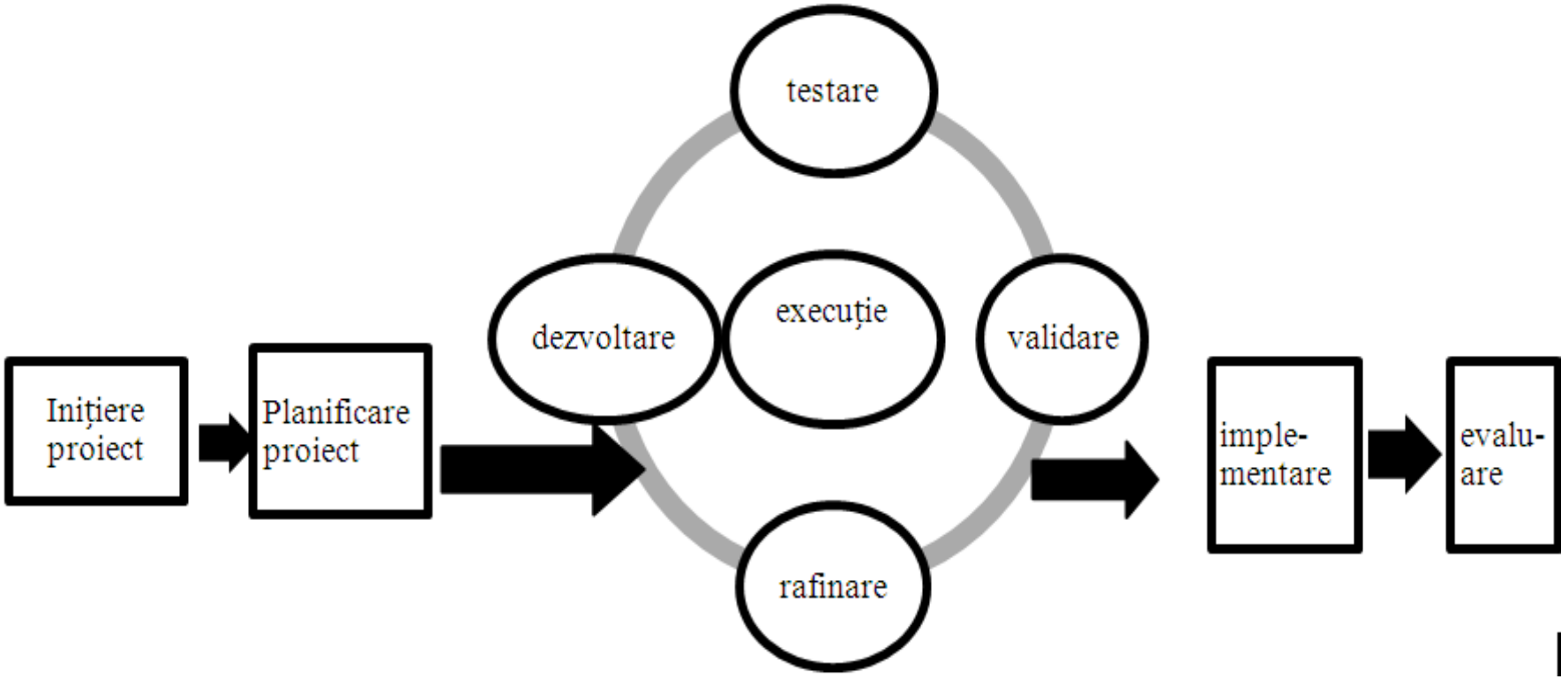
Beneficii

- Reducerea **duratei dezvoltarii** sistemului BI prin intelegerea sistemelor sursa
- **Acuratete** ridicata a **rezultatelor** BI
- **Transparenta crescuta** care le permite utilizatorilor si dezvoltatorilor sa isi dea seama ce informatii sunt disponibile

QPM- Qlikview Project Methodology

- Metodologie proprie QLIK, 2011
- Descrie activitatile legate de **managementul unui proiect** Qlikview si etapele de **realizarea a unei aplicatii BI**, inclusiv documente si livrabile
- Etape:
  1. Initierea proiectului,
  2. Planificare,
  3. Executie,
  4. Implementare
  5. Evaluare

QPM- Qlikview Project Methodology



Q1. Initierea proiectului

- **Definirea obiectivului initial** - corelat cu obiectivele de business;
- **Planificarea si bugetarea initiala activitati**:
  - Estimarea **duratei** proiectului si a perioadei de timp alocata fiecarei etape ;
  - Stabilirea **resurselor** necesare;
  - Definirea unui **plan initial cu activitatile** proiectului si cu perioada alocata fiecarei activitati;
  - Pregatirea **bugetului proiectului**;
  - Definirea **cerintelor initiale** cu urmatoarele activitati:
    i. Definirea cerintelor de business si a restrictiilor;
    ii. Identificarea cerintelor initiale legate de sursele de date

Q1. Initierea proiectului (cont)

iii. Întelegerea modului cum sunt create, stocate, transportate si raportate datele
iv. Stabilirea transformarilor necesare asupra datelor;
v. Identificarea cerintelor legate de integrarea datelor
vi. Realizarea unei mapari sursa-destinatie;
vii. Specificarea cerintelor infrastructurii
viii. Specificarea cerintelor de securitate (criptarea, autentificare si autorizarea accesului la date);
ix. Descrierea diferitelor solutii si utilizarea unui model SWOT pentru fiecare solutie. Identificarea **solutiei optime**.

Q2. Planificarea

A. **Planificarea managementului proiectului** cu urmatoarele activitati:
   - Actualizarea **cerintelor de business** si ierarhizarea lor
   - **Estimarea efortului necesar** pentru implementarea cerintelor de business.
   - Validarea **obiectivului** si a **scopului** proiectului;
   - Planificarea **etapelor de executie si implementare**;
   - Revizuirea **resurselor necesare** pentru urmatoarele etape si actualizarea planului de organizare a proiectului;
   - Alocarea **resurselor** la **roluri si responsabilitati**, alocarea rolurilor si a responsabilitatilor la fiecare **task**, pentru etapa de executie;
   - **Revizuirea bugetului**, tinând cont de ultimele modificari din planul proiectului;
   - **Analiza riscului**
   - Crearea **planului final** al proiectului.

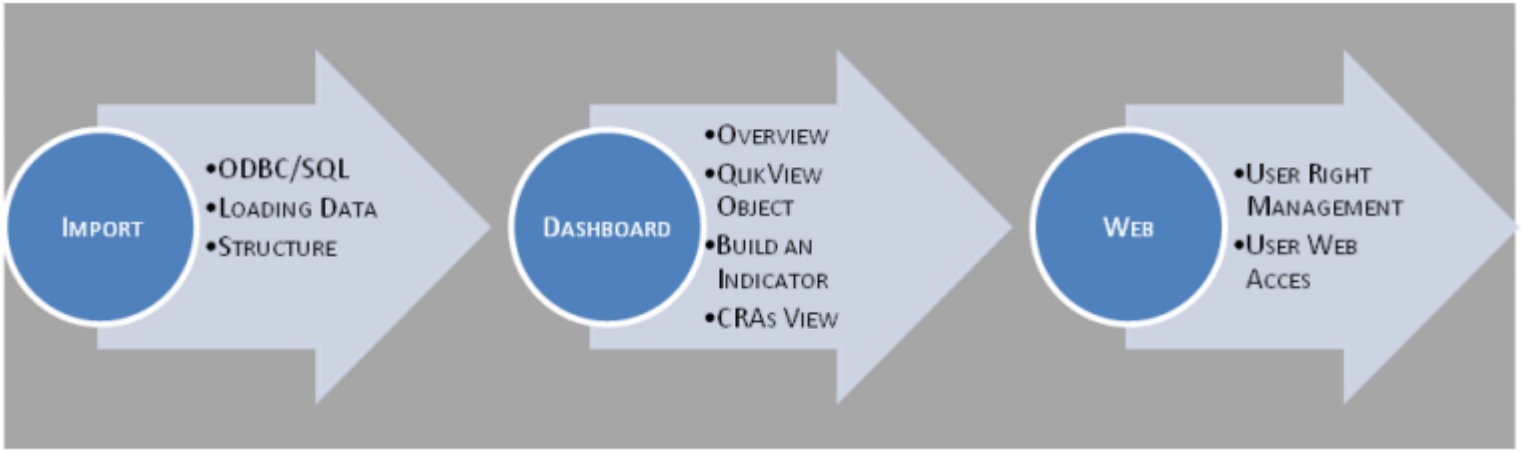
Q2. Planificarea (cont)

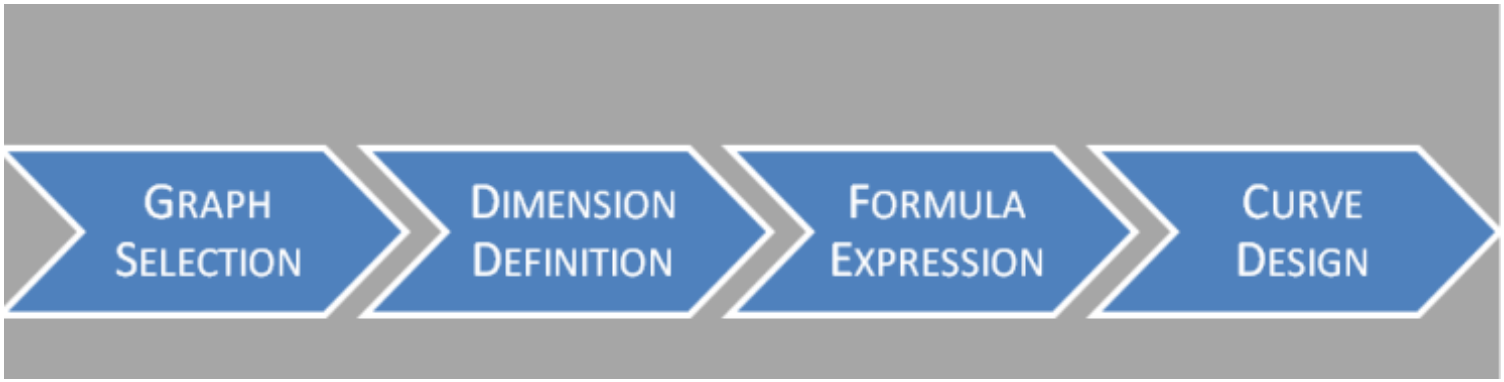
B. **Planificarea platformei Qlikview Enterprise cu urmatoarele activitati:**
   - Realizarea **modelului dimensional** initial
   - Definirea **cerintelor pentru ETL** (initiala si incrementala)
   - Definirea **arhitecturii aplicatiilor**
   - Identificarea **riscurilor asociate cu arhitectura** stabilita si evaluarea nivelului initial de risc

Q3. Executia – iterativa (3 sapt/iter)

- **Dezvoltarea**
  - dezvoltarea procesului de încarcare a datelor (configurarea conexiunilor, dezvoltarea scriptului de încarcare initiala a datelor);
  - crearea modelului de date (crearea fisierelor QVD);
  - dezvoltarea interfetei - abordare **DAR (Dashboards, Analysis, Reports)**
- **Testarea**
  - se va verifica daca sursele de date conectate sunt valide;
  - se va verifica corectitudinea expresiilor create;
  - se vor testa panourilor de bord pentru a verifica daca afiseaza indicatorii;
  - se vor testa diferite scenarii de business;
  - se va verifica daca a fost configurata corect securitatea aplicatiei.
- **Revizuirea si validarea de catre client**
- **Rafinarea solutiei**

Executia


Import
•ODBC/SQL
•Loading Data
•Structure
Dashboard
•Overview
•QlikView Object
•Build an Indicator
•CRAs View
Web
•User Right Management
•User Web Acces


Graph Selection
Dimension Definition
Formula Expression
Curve Design

Q4. Implementarea

- **Training-ul** utilizatorilor;
- Managementul **metadatelor**;
- Initierea procesului de **mentenanta**;
- **Migrarea** –mutarea aplicatiilor în productie;
- **Suport** pentru utilizatori.

Q5. Evaluarea

- evaluarea **aplicatiei BI** - mecanisme pentru îmbunatatirea solutiei BI
- evaluarea **managementului proiectului**,
- evaluarea **managementului riscurilor**,
- evaluarea **echipei de proiect**, a rolurilor si a responsabilitatilor asociate.

3. Advanced analytics

Analiza in-memory, big data

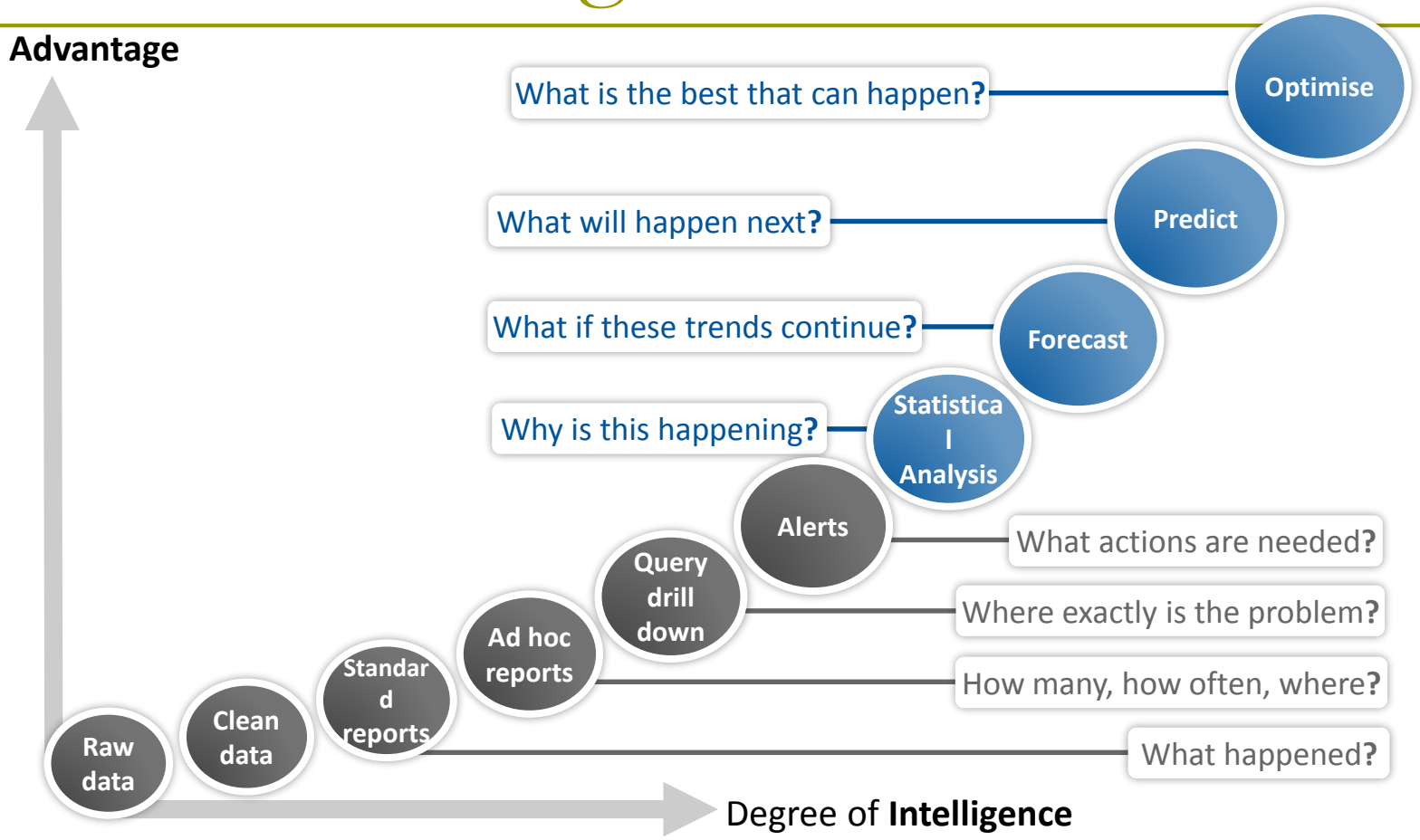
Tendintele majore in business analytics

- Advanced Analytics
- Mobile
- Cloud
- Social Media

Advanced Analytics

- O **categorie de metode de analiza** care pot conduce la schimbarea si imbunatatirea practicilor de business.
- Instrumentele de analiza traditionale – "BI este astazi ca si citirea unui ziar" = date istorice, ruleaza noaptea si produc date istorice
- **Advanced analytics** – previzionarea evenimentelor si comportamentelor **viitoare**, permitand realizarea de **analize what-if** pentru a prevedea efectele potentiale ale schimbarilor economice

Business Intelligence (vezi Cursul 1)



Business analytics

BA= business intelligence + advanced analytics

BA	Description	Examples
Descriptiv	Urmareste KPI pentru a intelege starea curenta a firmei (Ce se intampla?)	-enterprise reporting -Dashboards -Scorecards -OLAP -Data visualization -Simple statistical techniques
Predictiv	"Analizeaza trendul datelor pentru a evalua probabilitatea unor rezultate viitoare (Ce se va intampla?)	-Simulation -Advanced statistics -Data mining -Machine learning -Text analytics -Artificial Intelligence -Spatial machine learning
Prescriptiv	Utilizeaza performantele trecute pentru a genera recomandari despre rezolvarea unei situatii similare in viitor (Ce ar trebui sa fac?).	-Decision trees -Optimization and simulation algorithms -Fuzzy Rule-Based System -Switching Neural Networks

Advanced Analytics / Predictive Analytics

- Data Mining
- Regresii de date
- Simularea Monte Carlo

Exemple:

- Previzionarea comportamentului clientilor
- Segmentarea/clusterizarea clientilor
- Analiza cosului de cumparaturi
- Previziunea stocurilor

- Metode de analiza a textului ***- Text Analytics***
- Detectarea fraudelor – ***Anti-Fraud Analytics***
- ***Big data Analytics***

Aplicatii Data Mining

- **Predictive**:
  - Clasificare
  - Regresia
  - Detectarea deviatiilor
  - Filtrare colaborativa
- **Descriptive**:
  - Clustering
  - Descoperirea de reguli asociative
  - Descoperirea de modele secventiale

Amazon.com si NetFlix

Filtrarea colaborativa incearca sa prevada ce alte produse ar vrea clientul sa cumpere pe baza a ce a cumparat deja si a comportamentului altor cumparatori

Ce este Text Analytics?

• de exemplu, transformarea comentariilor nestructurate ale unui client in informatii utile, pentru a imbunatati actiunile companiei

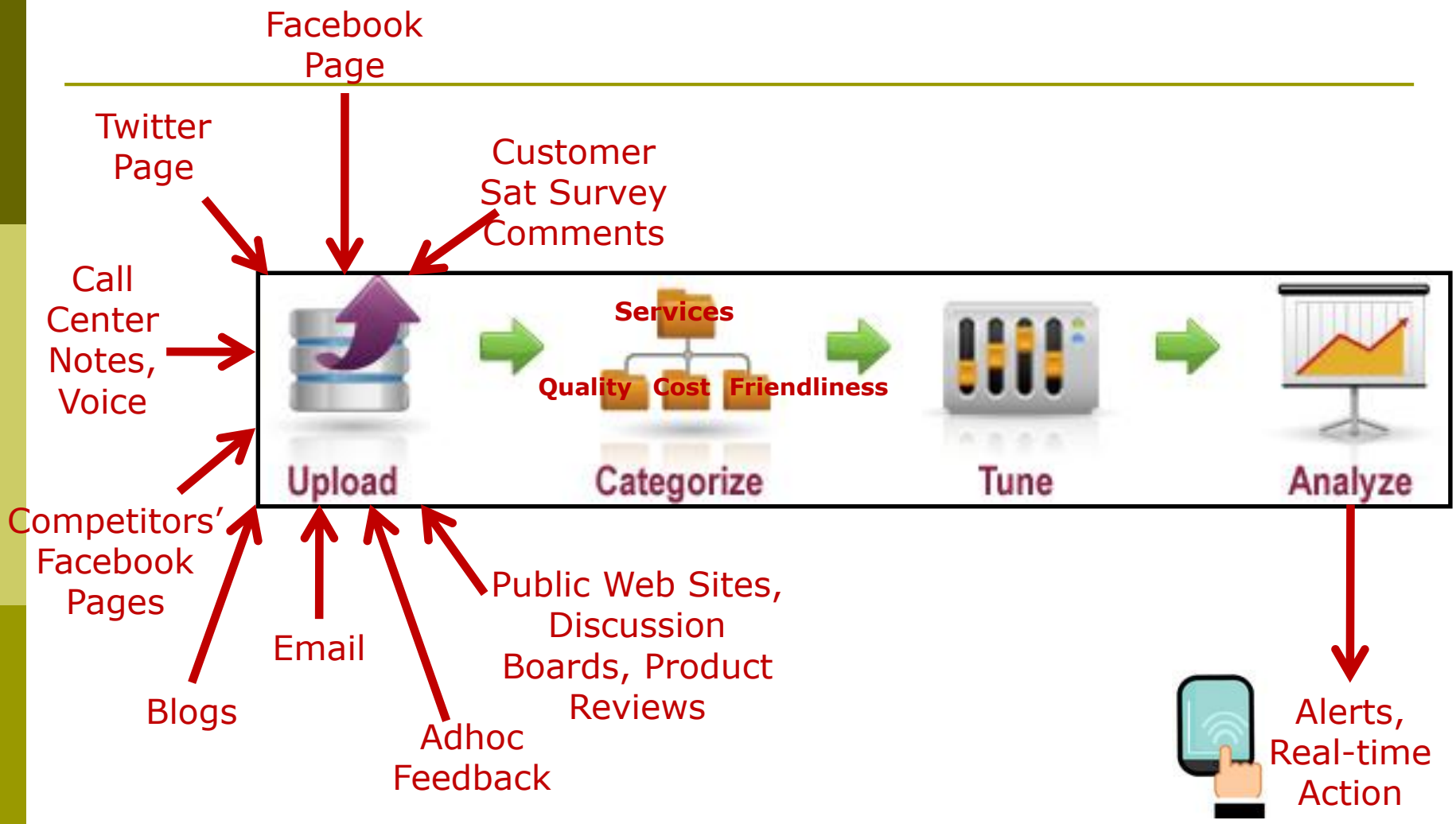
•***Wikipedia****:*

… un set de **instrumente lingvistice, statistice si tehnici de invatare automata** care modeleaza si structureaza continutul informational al surselor text pentru BI , analiza aprofundata a datelor si cercetare

•Tehnici de analiza bazata pe text:

- Social Analytics
- Sentiment Analysis
- Brand Identity
- Product & Brand Affinity
- Reputation Driven Online-Economy

Procesarea textului nestructurat



facebook 1 19 Search

Starbucks

Like

Food/Beverages

- Wall
- Info
- Starbucks Card
- International
- Ustream StarbucksLive
- Photos (8,737)
- Events
- Starbucks Jobs

More

About

Follow Starbucks on Twitter: http://twitter.com/Starbucks B... See More

20,327,043
people like this

Likes See All

- (RED)
- Starbucks Frappuccino
- Tazo Tea
- Fair Trade Certified
- Seattle's Best

Wall — Starbucks · Most Recent

Torbjørn Mcelroy
gett starbucks to norway :)
4 minutes ago

Jeanette A. Vollman
MM LOVE ME SOME ICE MOCHA!
19 minutes ago
2 people like this.

Renee Leo
I wonder if I could make more money and be happier if I worked in a Starbucks than at my job taking calls from irate citizens who have received traffic citations.
25 minutes ago
2 people like this.

Debi Jewell Renner I would love to work at Starbucks! Yummy!! Seriously though..I have the two best dream jobs ever!!! Thank you Jesus!!!
19 minutes ago

Kristen Pistro Neadow
I think some people need to learn how to fill the cup up and not with just ice.....

38 minutes ago
2 people like this.

Cautarea inteligenta

- **Facebook Graph Search**
  - A fost printre primii care au introdus procesarea limbajului natural ca functionalitate in algoritmul sau de cautare
  - Permite utilizatorilor sa formuleze solicitari in **limbaj natural**, de exemplu “persoane care iubesc masinile vechi” sau “prieteni interesati de rock” pentru a cauta in reteaua lor activa de pe Facebook sau “graful lor social”
  - Este un ***instrument puternic de mining*** pentru zona de ***marketing***, deoarece companiile pot afla siruri de corelatii care arata detalii mai profunde despre clientii lor

Factori : Modelele de procesare

- Data Mining
- Baze de date distribuite
- Baze de date in cluster
- Baze de date pe coloane
- In-memory Database Analytics
- In-database Analytics
- Real-time Data warehouses
- Procesare la sursa

In-database analytics

- Integrarea data analytics în funcționalitatea depozitelor de date.
- Există 3 tipuri:
  - traducerea unui model în cod SQL,
  - încărcarea bibliotecilor C sau C ++ în baza de date ca o funcție încorporată definită de utilizator (UDF)
  - bibliotecile out-of-process, de obicei scrise în C, C ++ sau Java și înregistrarea lor în baza de date ca UDF-uri încorporate într-o instrucțiune SQL.

“In-memory” BI

- Incarca setul de date in RAM – raspuns mult mai rapid
- SO pe **32 biti** puteau adresa doar **4 GB** de memorie RAM
- SO pe **64 biti** pot adresa pana la **1 terabyte** (TB) RAM
- Utilizeaza tehnici de **compresie** complexe si **stocarea pe coloane**
- Unele solutii **reduc/ elimina agregatele, cuburile**
- Reduce costurile IT si timpul de implementare al aplicatiilor BI

Solutia	Caracteristici	Exemple
In-memory OLAP	- Cub MOLAP incarcat in memorie - Lb MDX - modelare multidimensionala a datelor	IBM Cognos-Applix(TM1) Actuate BIRT
In-memory ROLAP	- metadate ROLAP incarcate in memorie - modelarea multidimensionala a datelor	MicroStrategy
O BD orientata pe coloane	- stocheaza datele intr-o BD orientata pe coloane - modelarea datelor - VizQL – lb declarativ	Tableau Software
In memory spreadsheet	- spreadsheet incarcat in memorie - Lb DAX (data analysis expressions)	Microsoft PowerPivot
In memory "associative" data model	- stocheaza datele intr-un model "asociativ" incarcat in memorie - toate jonctiunile si calculele se fac in timp real - scripturi pt incarcarea si transformarea datelor	QlikView
Abordare hibrida cu tehnici de compresie	BDR + BD orientata pe coloane	Oracle Exalytics In-memory (include Essbase, in –memory TimesTen datebase); SAP HANA
disk +in-memory	-MOLAP - MDX- stocheaza agregatele si datele atomice pe disc -Tabular -lb DAX – stocheaza datele atomice in memorie	SQL Server

Qlikview

- utilizeaza un **model de date “in-memory” -** stocheaza toate datele in RAM, deci timp de raspuns mai mici
- utilizeaza **algoritmi de compresie** complecsi
  - datele sunt comprimate la 10% din dimensiunea lor originala atunci cand sint incarcate in documentul QlikView
- utilizeaza diferite **surse de date**:
  - baze de date (conexiune prin ODBC, OLEDB), fisiere (Excel, CSV, HTML, XML, etc.)
  - exista de asemenea, diferiti **conectori** la aplicatiile SAP, Salesforce, retele sociale (ex. Twitter)

Surse de date suportate ca intrari

A-H	I-O	P-Z
Actian Vectorwise	IBM DB2	ParAccel
Amazon EC2	IBM Netezza	ParStream
Amazon Redshift	IBM (Lotus) Notes	PostgreSQL
Aster Data nCluster	Infor Lawson	Progress OpenEdge
Cloudera Hadoop Hive	Intuit QuickBooks	Sage 500
Cloudera Impala	Informatica Powercenter	Salesforce
CSV	MapR	SAP
DataStax	MicroStrategy	SAP HANA
Epicor Scala	Microsoft Access	SAP NetWeaver Business Warehouse

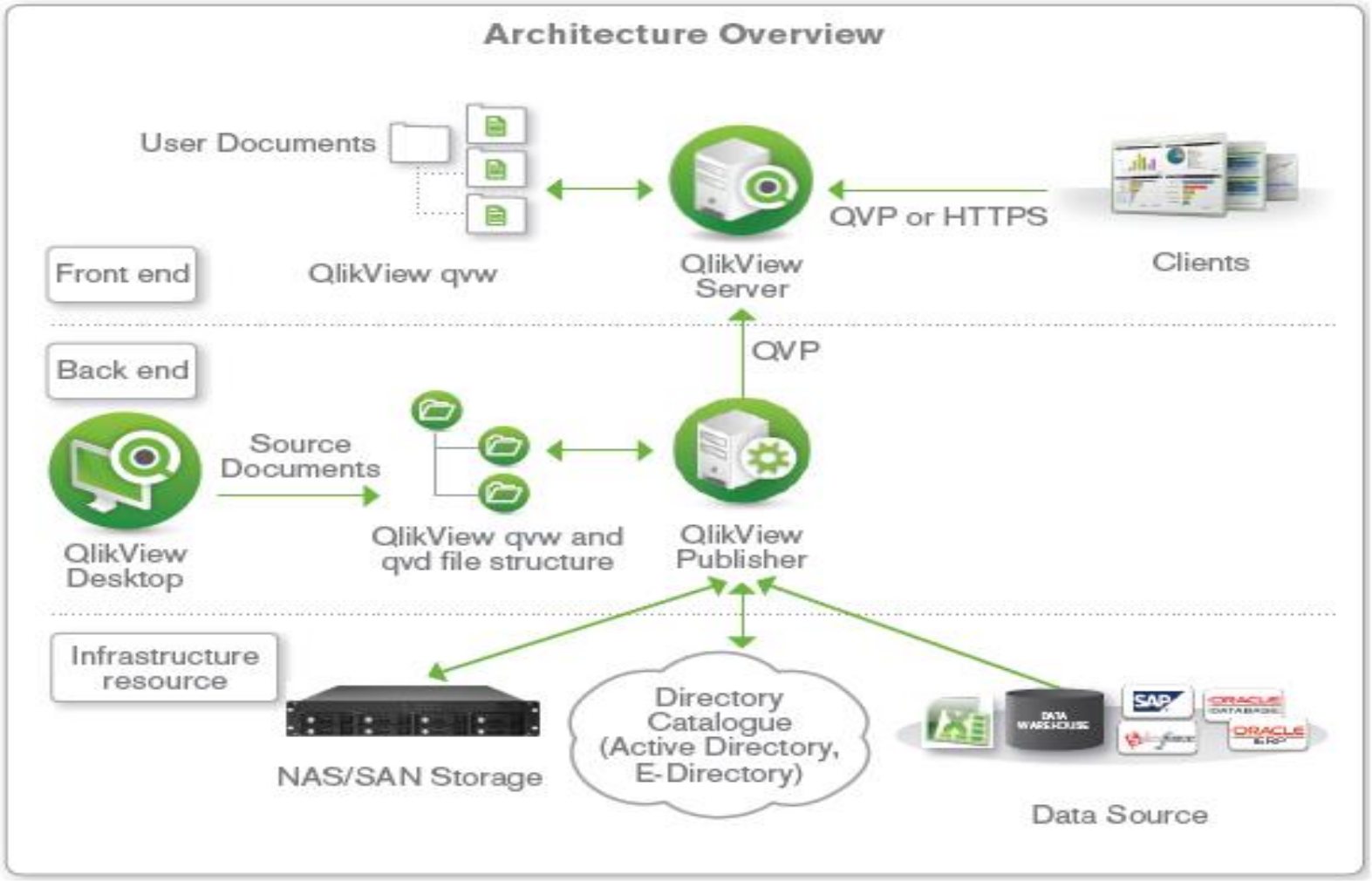
Surse de date suportate ca intrari

EMC Greenplum	Microsoft Dynamics NAV (Navision)	Siebel
Firebird	Microsoft Excel	Sybase ASE
Google BigQuery	Microsoft SharePoint	SybaseIQ
Hortonworks Hadoop Hive	Microsoft SQL Server	Teradata
HP Vertica	MySQL	Web pages
	OData	XML
	ODBC	
	Oracle	
	Oracle Hyperion	
	Oracle JD Edwards	
	Oracle Peoplesoft	

Iesiri de date

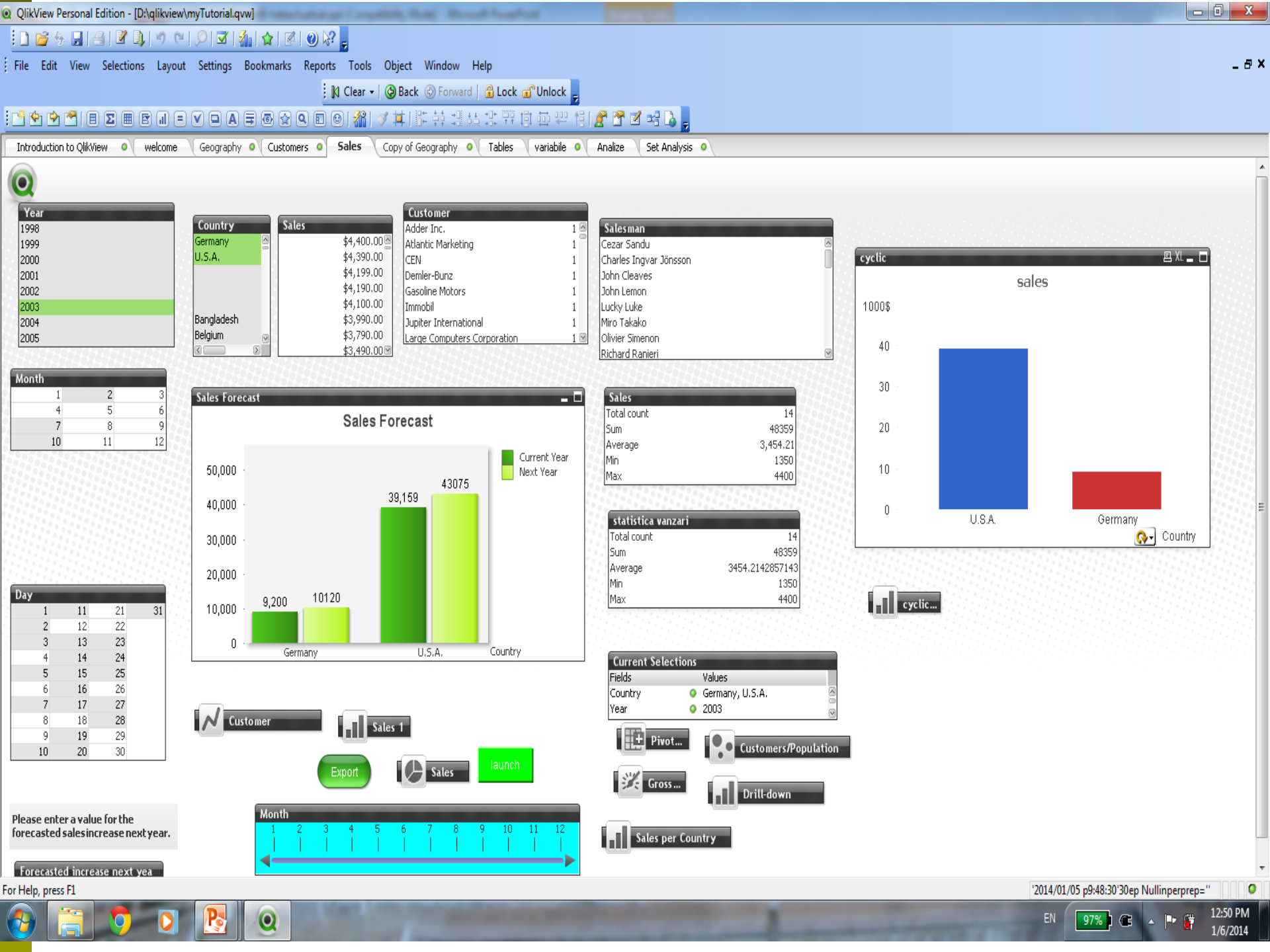
- QlikView poate exporta in ***Excel, fisiere imagine, Pdf***
- Este capabil sa preia datele din BD si sa exporte rezultatul in Excel
- Poate salva rezultatul interogarilor in Excel, Pdf si PowerPoint
- Poate sa faca **jonctiuni** pe diferite tipuri de tabele si sa le exporte in Excel, Word, Pdf
- Suporta operatiile de ***drill-up***, ***drill-down***, ***slicing*** si ***dicing***
- Ofera rapoarte /tablouri de bord **in timp real**, depinde de rata de refresh stabilita
- Rapoartele/ tablourile de bord pot fi **afisate pe Internet/Intranet** cu date reale sau cu refresh la anumite momente
- Poate sa salveze rezultatul interogarilor pentru **utilizare ulterioara**

Arhitectura QlikView

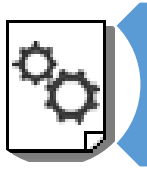


Qlikview

- **script de incarcare –**poate fi utilizat pt a extrage, transforma si **incarca datele** in modelul de date sau **pt a stoca modelul (inclusiv datele)** pe disc in **fisiere intermediare (QVD).**
- datele sunt **stocate la nivel de detaliu**, toate agregatele se realizeaza “on the fly”, la cerere
- Are un motor de interfata care intretine automat asocierile intre date
- **selectiile** facute de utilizator se propaga in cascada prin tot modelul de date.
  - Orice selectie facuta in documentul Qlikview este automat aplicata pe intregul model de date.
- aplicatiile QlikView pot rula pe clienti/SO multiple


QlikView Personal Edition - [D:\qlikview\myTutorial.qvw]
File Edit View Selections Layout Settings Bookmarks Reports Tools Object Window Help
Clear Back Forward Lock Unlock
Introduction to QlikView welcome Geography Customers Sales Copy of Geography Tables variabile Analize Set Analysis
Year
1998
1999
2000
2001
2002
2003
2004
2005
Country
Germany
U.S.A.
Bangladesh
Belgium
Sales
$4,400.00
$4,390.00
$4,199.00
$4,190.00
$4,100.00
$3,990.00
$3,790.00
$3,490.00
Customer
Adder Inc. 1
Atlantic Marketing 1
CEN 1
Demler-Bunz 1
Gasoline Motors 1
Immobil 1
Jupiter International 1
Large Computers Corporation 1
Salesman
Cezar Sandu
Charles Ingvar Jönsson
John Cleaves
John Lemon
Lucky Luke
Miro Takako
Olivier Simenon
Richard Ranieri
cyclic
sales
1000$
40
30
20
10
0
U.S.A.
Germany
Country
Month
1 2 3
4 5 6
7 8 9
10 11 12
Sales Forecast
Sales Forecast
50,000
40,000
30,000
20,000
10,000
0
9,200
10120
39,159
43075
Current Year
Next Year
Germany
U.S.A.
Country
Sales
Total count 14
Sum 48359
Average 3,454.21
Min 1350
Max 4400
statistica vanzari
Total count 14
Sum 48359
Average 3454.2142857143
Min 1350
Max 4400
cyclic...
Day
1 11 21 31
2 12 22
3 13 23
4 14 24
5 15 25
6 16 26
7 17 27
8 18 28
9 19 29
10 20 30
Current Selections
Fields Values
Country Germany, U.S.A.
Year 2003
Customer
Sales 1
Export
Sales
launch
Pivot...
Customers/Population
Gross...
Drill-down
Sales per Country
Month
1 2 3 4 5 6 7 8 9 10 11 12
Please enter a value for the forecasted salesincrease next year.
Forecasted increase next yea
For Help, press F1
'2014/01/05 p9:48:30'30ep Nullinperprep=''
EN
97%
12:50 PM
1/6/2014

Implementare BI traditionala



Implementari indelungate, luni sau chiar ani

Necesar crescut de servicii profesionale

Costuri ridicate

Accent pe implementare IT

Implementare QlikView

Implementari rapide - zile sau saptamani

Necesar minim de servicii profesionale

Costuri reduse

Accent pe modelul de afaceri

Alte beneficii

- **Disponibilitate ”in cloud”** - Prin intermediul Amazon’s Elastic Compute Cloud (EC2) Web service
- **QlikView pentru smartphone și telefoane mobile -** Clienți dedicați Java Mobile și iPhone si Android; Qlik Sense app
- **Simplu și utilizabil –** Caracteristici de căutare și vizualizare în mod grafic îmbunătățite
- **QlikView Personal Edition –** Disponibil gratuit pentru dezvoltatori

Limitari QlikView

- Produsul a fost proiectat pentru **aplicatii BI la nivel de GByte** (10-20 GB de date compresate), datele fiind incarcate in intregime in memoria RAM
- Aceasta limita a fost depasita in Qlik Sense
- Ofera o **pregatire limitata a datelor** – limbajul de scripting pentru incarcare si integrare *nu ofera capacitati ETL avansate*
- **Nu ofera o calitate inalta a aspectului rapoartelor**, fiind creat in special pentru analiza interactiva. Pentru a crea un design mai bun e necesara utilizarea *macrourilor*

Big Data Analytics

- **Cat de multe date?**
  - Google proceaza peste 20 PB pe zi
  - Facebook are 2.5 PB de date utilizator + 15 TB/zi
  - eBay are 6.5 PB de date utilizator + 50 TB/zi
- **Ce tipuri de date?**
  - Date relationale
  - Text (Web)
  - Date semistructurate (XML)
  - Date sub forma de graf
    - Social Network, Semantic Web
  - Streaming Data
    - Datele se pot scana o singura data

Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Big data : Volum, Viteza si Varietate

- **Volum:** companiile se confrunta cu tera sau chiar petabytes de informatii.
  - 350 bilioane de citiri de contor pentru a **previziona** consumul de energie
- **Viteza:** Existe procese care sunt **sensibile la timp**, ex detectarea **fraudelor**
  - Parcurgerea a 5 milioane de tranzactii de vanzare zilnic pentru a detecta fraude
  - Analiza 500 milioane apeluri zilnice de la clienti pentru a prevedea mai rapid **nemultumirile** clientilor
- **Varietate:** text, date de la senzori , audio, video, click streams, fisiere log
  - Filme provenite de la camerele de supraveghere
  - Exploatarea cresterii de 80% a datelor sub forma de imagini, video si documente pentru cresterea satisfactiei clientilor

Tipuri de instrumente folosite de obicei in Big Data

- Unde are loc procesarea?
  - Distribuita
- Unde sunt stocate datele?
  - Stocare distribuita (ex: Amazon s3)
- Care este modelul de procesare?
  - Procesare distribuita (Map Reduce)
- Cum sunt stocate si indexate datele?
  - Schema performanta, indiferent de baza de date
- Ce operatii se realizeaza pe date?
  - Procesare analitica/Semantica (Ex. RDF/OWL)

Apache Hadoop

- **2008** – mai intai Yahoo, Ebay sau Facebook
- platforma **open source**
- procesare distribuita pe clustere de servere
- standard “de facto”
- **Java** based framework
- modele de procesare paralela
- cod in orice limbaj contemporan (API)
- SCALABILITATE, ROBUST
- **Hadoop Distributed File System (HDFS):** stocare in cluster
- **MapReduce:** motor ce procesare paralela - management distribuit al resurselor

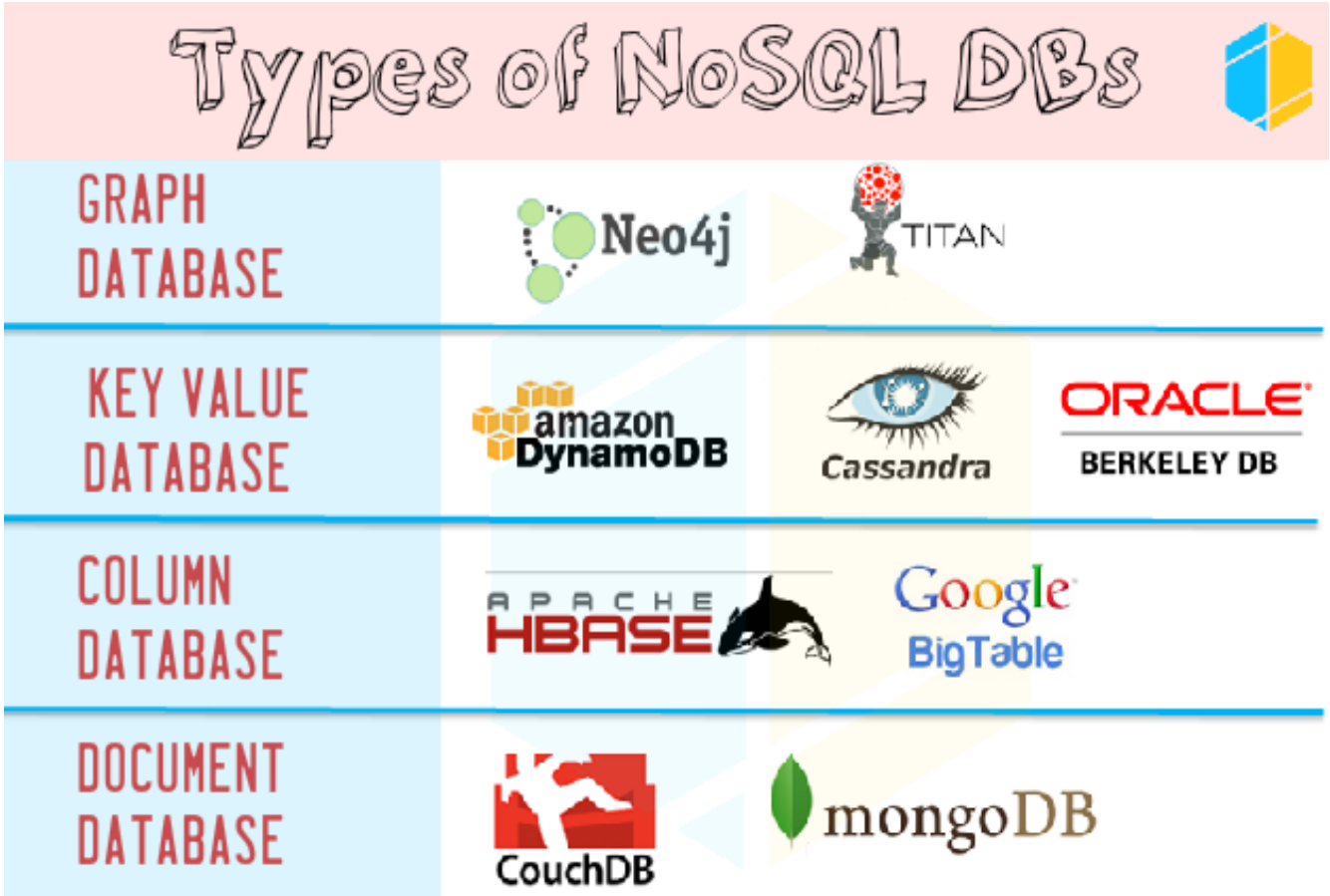
Tehnologii Big data

- Apache Hadoop
  - Ofera suport aplicatiilor distribuite orientate pe date
  - Permite aplicatiilor sa lucreze cu mii de calculatoare care proceseaza independent, petabytes de date
- MapReduce
  - **Nucleul Hadoop**- o platforma si un model de programare cu scalabilitate masiva care poate procesa seturi imense de date pe sute sau mii de servere (noduri) din clusterul Hadoop.
  - **Map()**: converteste seturile de date de pe un nod in inregistrari (perechi cheie-valoare)
  - **Reduce()**: combina inregistrarile de la noduri in setul de date cerut
- NoSQL
  - **MongoDB** (Apache): o baza de date orientata pe documente care foloseste structuri tip JSON in locul structurilor de date relationale.

Baze de date NoSQL

- pentru aplicații în cloud,proiectate să rezolve problemele de scalabilitate, performanță, modelare și distribuire limitată a datelor din bazele de date relaționale.
- nu impun o anumită schemă (structură a datelor), au un API simplu, sunt ”eventual consistente” și pot gestiona o cantitate foarte mare de date
- **Teorema CAP** (Eric Brewer) – orice sistem distribuit de management al datelor poate sa indeplineasca maxim doua din urmatoarele trei proprietati:
  - **C: Consistenta** - toate nodurile sistemului informatic stocheaza aceleasi date
  - **A: Availability / Disponibilitate** - orice cerere va primi un raspuns
  - **P: Partitionare** - sistemul continua sa functioneze in conditii de partitionare a retelei
- **Arhitectura distribuita**
- **Toleranta la defecte**.

Tipuri de BD noSQL



Avantaje

- **Nu este necesar ETL-** suporta stocarea datelor "ca atare, iar transmiterea de date catre alte sisteme se poate face sub forma de document XML, JSON sau obiect binar
- **Suport pentru structuri de date multiple**
- **Lucrul cu medii distribuite**
- **Portabilitatea -** NOSQL utilizeaza modelul DHT (Distributed Hash Table), manipularea datelor obiect se realizeaza prin cheia primara a obiectului.