Data in Online Business

D. Pejčoch

(eBay UK Ltd. + KIZI VŠE Praha)

Outline

- What is online business
- Examples of business models
- Typical processes in online business
- Big volume of processed data
- Specific problems with data
- What is Big Data?
- Potential sources of Big Data
- Specific issues in data quality

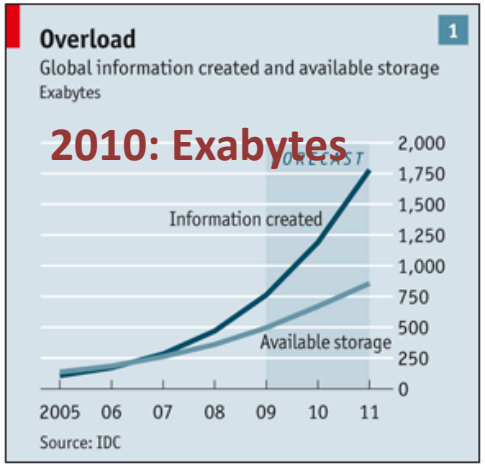
What is online (.com) business?

Examples of Business Models

- **eShop:** revenue = buying price – selling price = insertion fee + final value fee + feature fee + subscription fee + ...
- **Google:** 98% of revenue from selling ad space
- **Facebook and others:** advertising, payment revenues, ... and ...?
- **Zynga, Geewa, ...:** virtual goods, advertising
- **Instagram (FB), ...:** who knows? ... perhaps they sell customer data?
- **LinkedIn:** Freemium business model (InMails, Profile Stats Pro, ...)
- **Slideshare (LI):** Freemium as well
- **Foursquare:** still has no clue

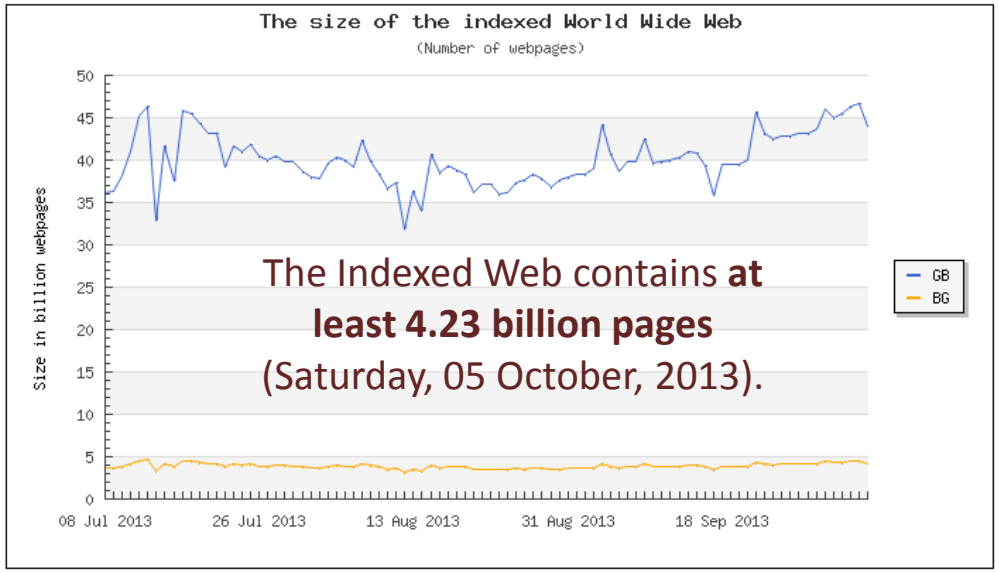
Comparison of processed volume of data

http://www.economist.com/node/15557443



http://www.worldwidewebsize.com/

Last Month | Last Three Months | Last Year | Last Two Years

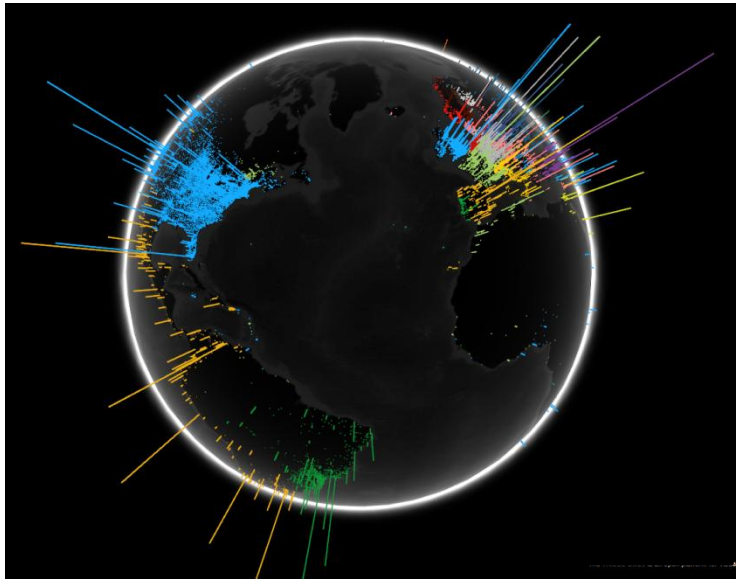


The Indexed Web contains **at least 4.23 billion pages** (Saturday, 05 October, 2013).

GB = Sorted on Google and Bing
BG = Sorted on Bing and Google

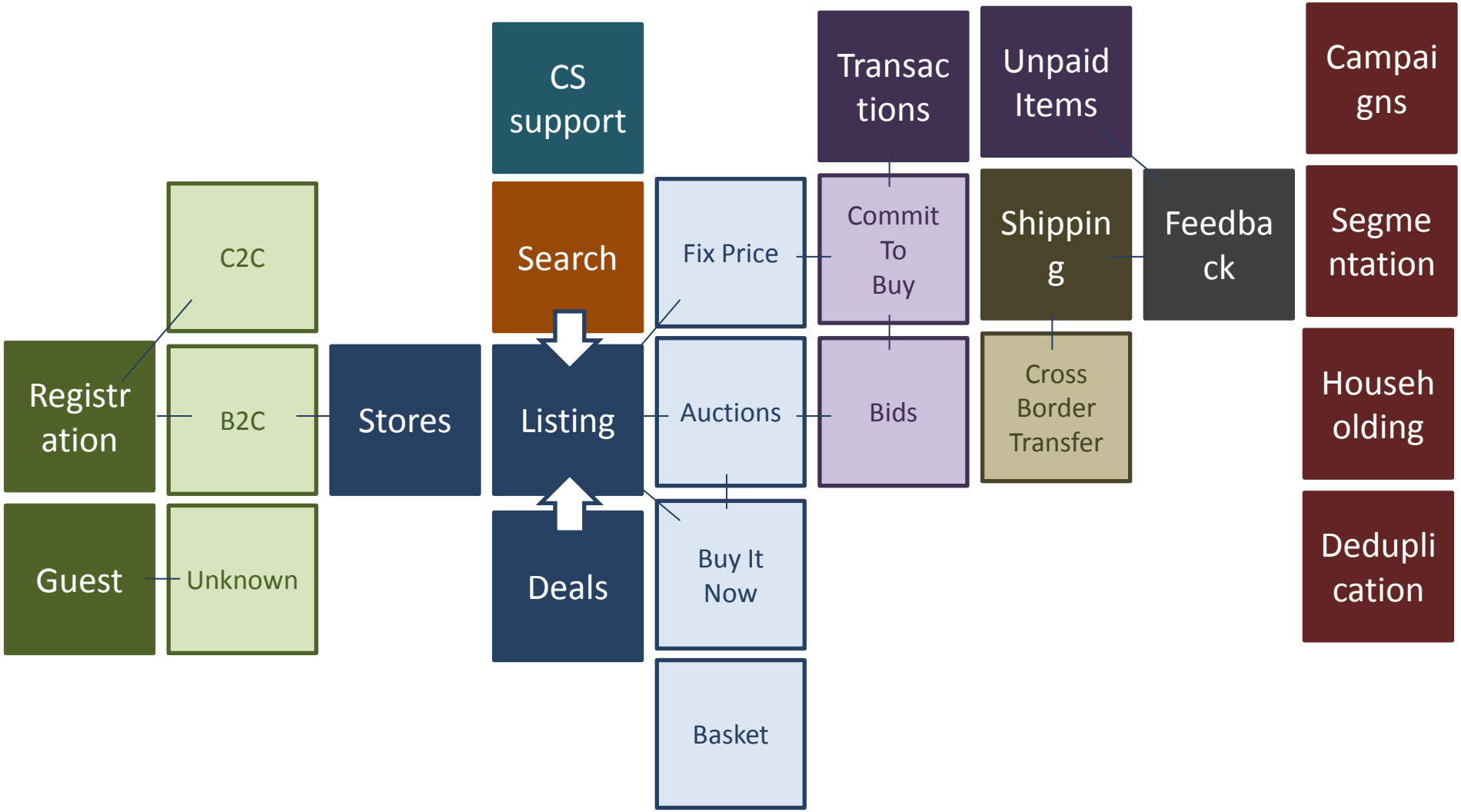
Specific problems with volume of data

- Huge data centers
- High availability
- Millions rows per day
- Even robust RDBMS are not enough
- Limited DB space and resources assigned to single analyst
- Extreme effort in optimalization of SQL queries
- Long queues of queries waiting for execution
- Long time to process single query
- Logical + physical partitioning in data
- Tables with 1% samples



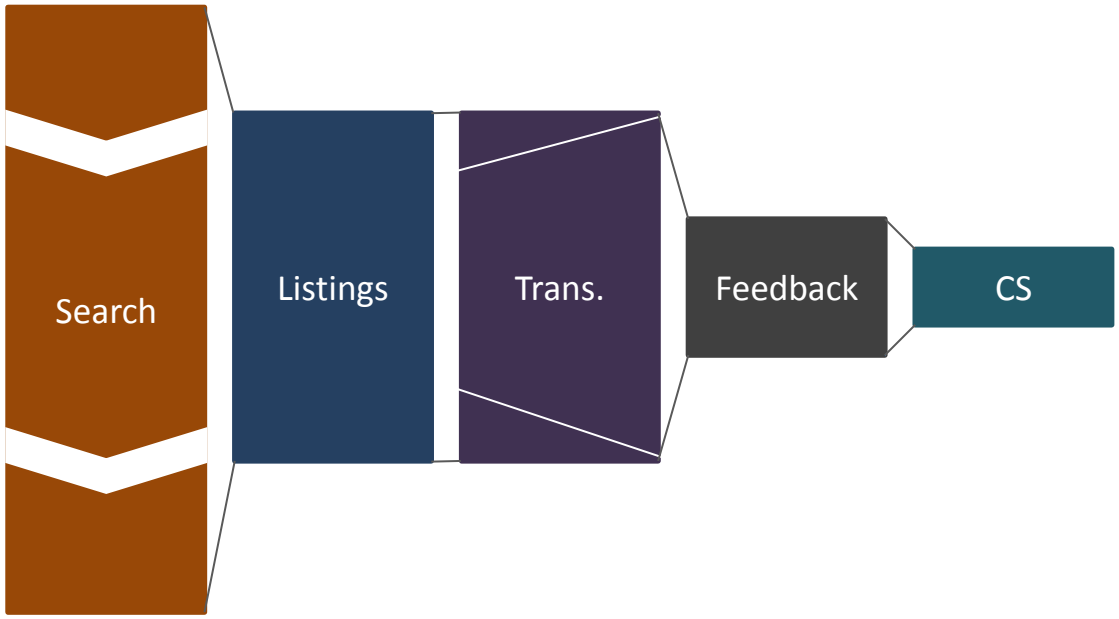
http://data-arts.appspot.com/globe-search/

Elementary Table of Online Business: e-shop

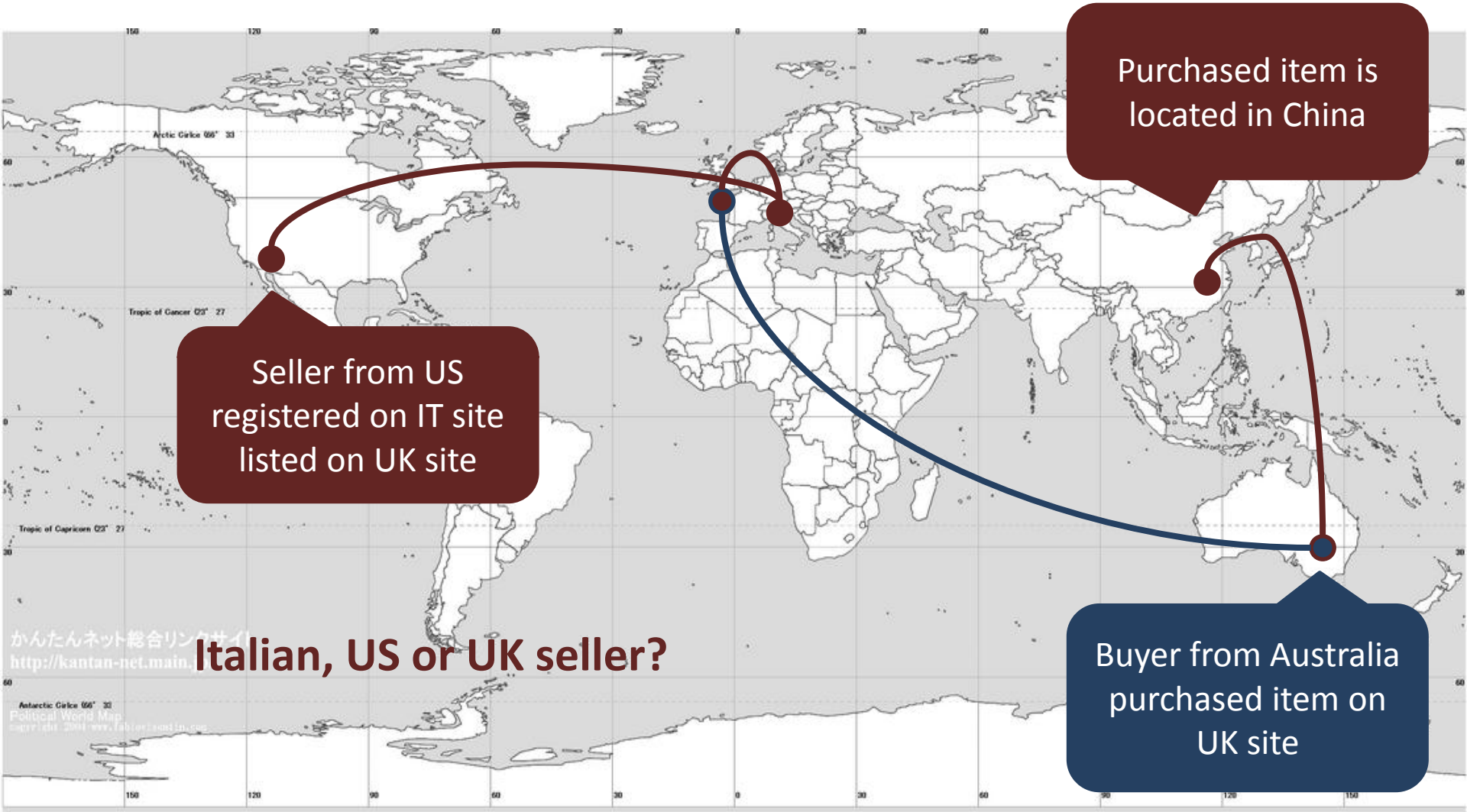


Classification of Processed Data

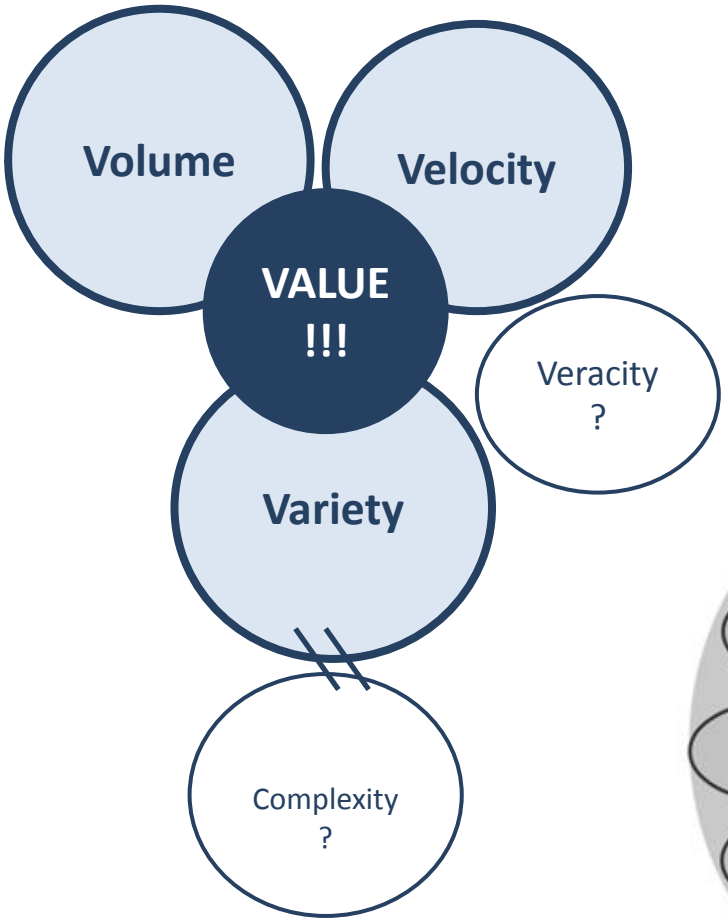
- Metadata:
  - Technical: e.g. description of huge data models
  - (Business: labels, description in reports)
- Master Data:
  - Customers, their hierarchies and interactions (e.g. Facebook)
  - Product Categories
  - Stores
- Transactional Data:
  - Search
  - Listings
  - Bids
  - Transactions
  - Revenue
  - Shipping
  - Payment
  - Feedback
  - Campaigns
  - Promos
  - Deals
  - …
- External Data + LOVs



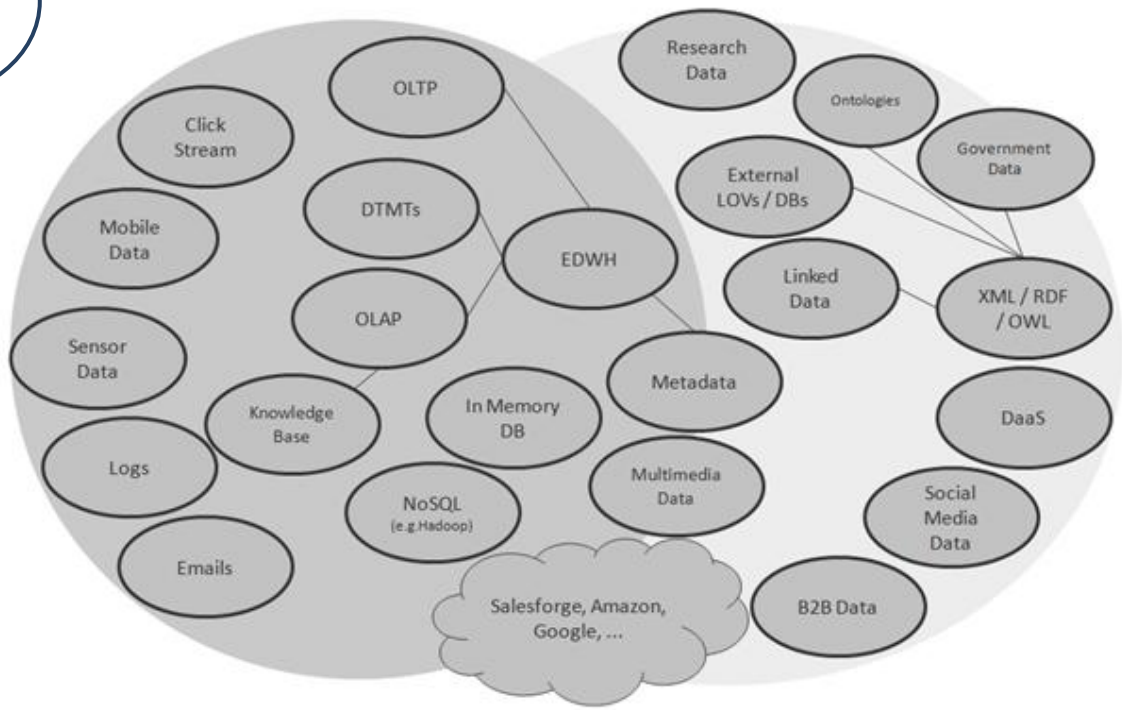
Global World of Internet Business: eBay Transaction



What is Big Data?



- Data which is impossible or not effective to process using common RDBMS
- Gartner‘s Vs
- Big Data = Hadoop
- Processing of large unstructured data
- Different sources
- Big Data = Small Data + Big Garbage?



Big Data and online business

- **Hadoop** has been developed within online business (Google)
- **Typical use**: search indexation
- **Secondary use**: when RDBMS is not sufficient, when faster delivery is critical, when aditional knowledge is needed, ...
- **Big Data Quality**: Just a New Buzzword or Serious Topic?
  - Quality within Hadoop => re-active DQM => spaghetti code, weakness of Hive SQL, DQM decentralization ...
  - Quality of sources => pro-active DQM => missing metadata, lost control, ...
  - Big Data as reference source
  - Hadoop as DQM platform

Big Data
is only another kind of data
under
global Governance
initiative
!!!

http://markosun.wordpress.com/2012/10/21/inside-googles-giant-data-centers/

My Tips for Potential Sources of Big Data

- **Social Networks** => Word-of-mouth (e.g. testing change of fees), Loyalty program
- **Registers of Debtors** => UPI prevention
- **Shipping data (using RFID)** => Shipping optimization
- **Weather information** => predicting BYR / SLR behavior => better targeting campaigns and promo actions
- **Actual news** => explaining production metrics anomalies
- **External LOVs** => Demographic and behavioral informations => better set up of campaigns and Next Best Offers, better segmentation based on client profiles
- **Informations about Competitors** (results of Competitive Intelligence)
- **External Metadata** => better validation process
- **Analysis of plaintext attributes**
  - call center: needs, validation, process maturity improvement, new attributes
  - items description: avoid IAD Low DSRs
- **Product information** (e.g. EAN) => better analysis of product (e.g. CBT, product specific MOTs)

Specifics of Data Quality / Governance

- **Not standardized data**: fields in item description / title, tweets, fbk statements, ...)
- A lot of **not structuralized** data => specific problems with quality
- **Hadoop accuracy**
- Product catalogue on basis of product categories (**small level of granularity**)
- **Deduplication** of users (guests vs. standard users, different accounts, purchases without registration, ...) ... thanks God for cookies! FOAF: thanks for emails
- Customer >= User = Account
- Importance of **metadata** and management of **relevant knowledge**
- **Household** identification (no care => no verification)
- Not **online MDM:** performance (too complex)
- **Global environment** (mix of Locales)
- **Total Costs of Information** <= too big data
- **Regulations:** SOX, Data Quality Act, ... no Basel, no Solvency, ...
- **Governance:** not all data under (direct) control
- **Stewardship:** complex knowledge => data scientists?