Machine Learning and Cloud Computing

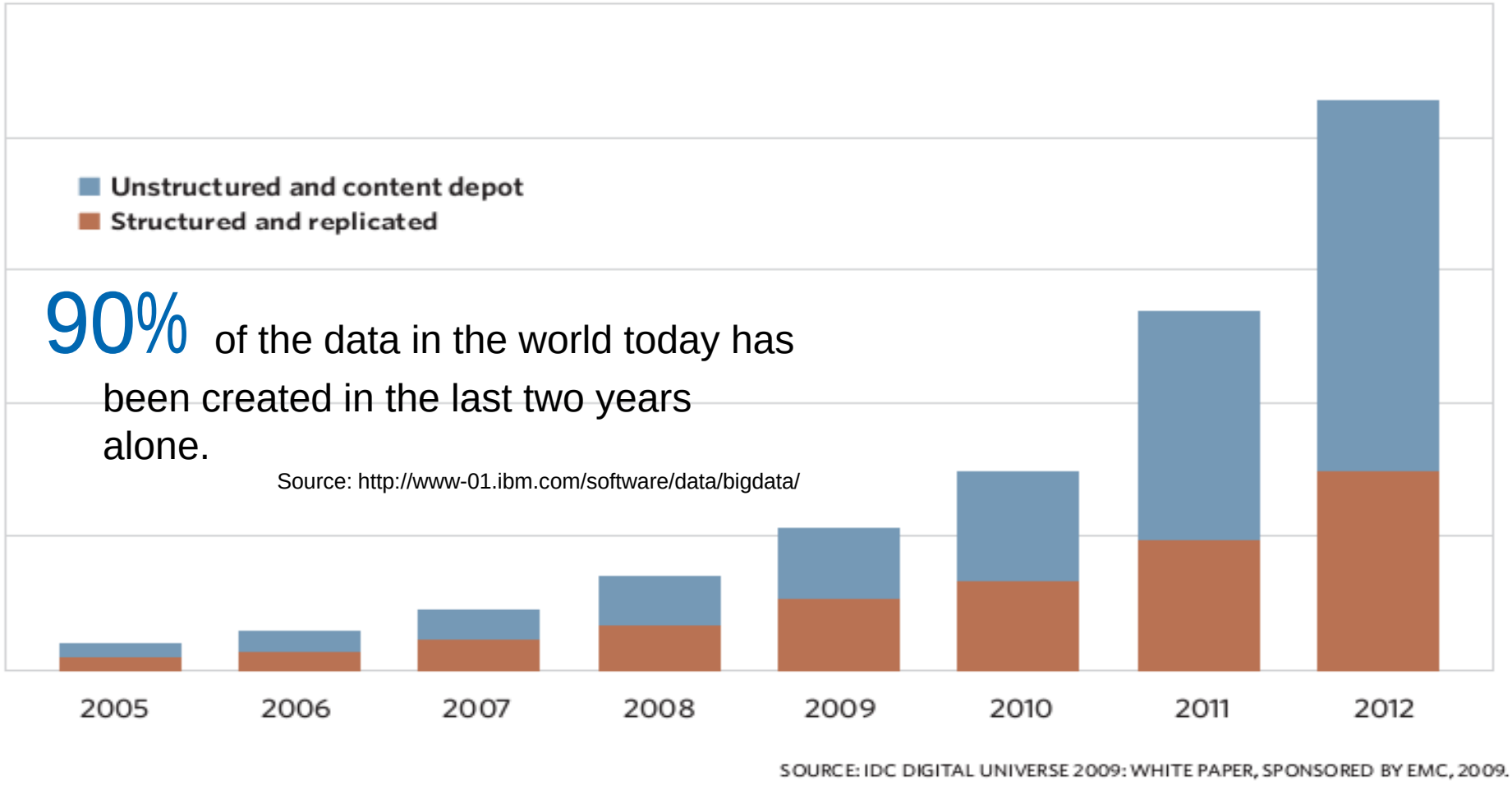
trends, issues, solutions

Daniel Pop

Future plans // Tools and methods

- Develop software package(s)/libraries for scalable, intelligent algorithms for distributed environments (cloud, grid, cluster) and validate on real areas: (...)

Big Data



90% of the data in the world today has been created in the last two years alone.

Source: http://www-01.ibm.com/software/data/bigdata/

SOURCE: IDC DIGITAL UNIVERSE 2009: WHITE PAPER, SPONSORED BY EMC, 2009.

Big Data Processing

In ‘industrial revolution’ terms, we are in the pre-industrial era of artisanship that preceeded mass production.

It is the equivalent of needing to engage an expert blacksmith to forge the forks and spoons for our dinner table.

Ben Werther, co-founder Platfora

4 Vs of BigData

- Volume:
  - 12TB of tweets / day => product sentiment analysis
- Velocity:
  - fraud detection
  - predict customer churn faster (anlyse 500 million daily calls in real-time)
- Variety:
  - exploit 80% growth of un&semi-structured data for customer satisfaction
- Variability / Veracity:
  - variance in meaning (1 in 3 business leader don‘t trust the information they use to make decisions)

Brian Hopkins (Forrester)

Paradigm shift drivers

- Expressing ML-DM algorithms in SQL is complex and hard to maintain
- Large-scale installations of parallel relational databases is expensive
- New type of data: un/semi-structured
- Existing ML tools – does not offer good support for processing large sets of data

New technologies

- NoSQL data stores – distributed data storage solutions
- MapReduce – distributed parallel processing environment

Distributed solutions

1. ML environments in the cloud
2. Plugins for ML tools
3. Distributed ML libraries
4. ML systems
5. Software as a Service providers for ML

Daniel POP - Machine Learning and Cloud Computing: Survey of distributed and SaaS solutions, Technical Report Institute e-Austria Timisoara, 2012

Machine Learning environments in the cloud

Create a cluster in the cloud and bootstrapping it with statistics tools

- Cloudnumbers – R, Octave, Maple / Amazon EC2
- CloudStat – R
- Opani – R, Python / Rackspace

- Access to low-level functionalities of statistics tools from console or graphical interfaces
- Support for visualization, data connectors, cluster monitoring

Plugins for Machine Learning tools

Augment statistics tools (R, Python) with plugins that allow users to create a cluster and run ML jobs on it

- RHIPE – R + Hadoop & HDFS
- Snow & variants – R + sockets / MPI / PVM
- Segue for R – R + Amazon Elastic Map Reduce
- Anaconda – Python

- Target users: mathematicians, statisticians, programmers, data analysts, machine learning
- Re-use existing private e-infrastructure

Distributed ML libraries

Collection of parallelized implementations of ML algorithms for distributed environments (Hadoop, Dryad, MPI etc.)

Name	Platform	Licensing	Language	Activity
Mahout	Hadoop	Apache 2	Java	High
GraphLab	MPI / Hadoop	Apache 2	C++	High
DryadLINQ	Dryad	Commercial	.NET	Low
Jubatus	ZooKeeper	LGPL 2	C++	Medium
NIMBLE	Hadoop	?	Java	Low
SystemML	Hadoop	?	DML	Low

Distributed ML libraries – cont.

- Offer access to out-of-the-box, optimized implementations of ML algorithms
- Offer a framework for custom implementations of ML algorithms
- Target users: developers, architects etc.

Machine Learning systems

Products that need to be installed on private data centres (or in the cloud) and offers high performance data mining and analysis

- Kitenga Analytics
- Pentaho Business Analytics
- Platfora
- Skytree Server
- Wibidata

Machine Learning systems

- Complex systems
- Deployable on-premise
- Rich set of graphical tools
- Expose a limited set of ML-DM algorithms (with few exceptions)
- Hadoop-based

- Target users: business users

SaaS providers of ML

PaaS/SaaS solutions that allow clients to access ML algorithms via lightweight (RESTful) Web services

- BigML
- BitYota
- Precog
- Google Prediction API
- EigenDog
- Metamarkets
- Myrrix
- Prior Knowledge Veritable API
- Predictobot

- Predictive modelling
- Lack of customization
- Target: systems (recommendation, analysis), business users

Text mining as SaaS

PaaS/SaaS solutions that allow clients to access NLP and text mining algorithms via Web services

- Alchemy API
- Nathan App
- Text Processing
- Yahoo! Content Analysis Web Service

- Sentiment analysis, entity extraction & recognition, semantic learning, document summarization and clustering etc.
- Social media, Semantic web

Why?

Issues with current solutions

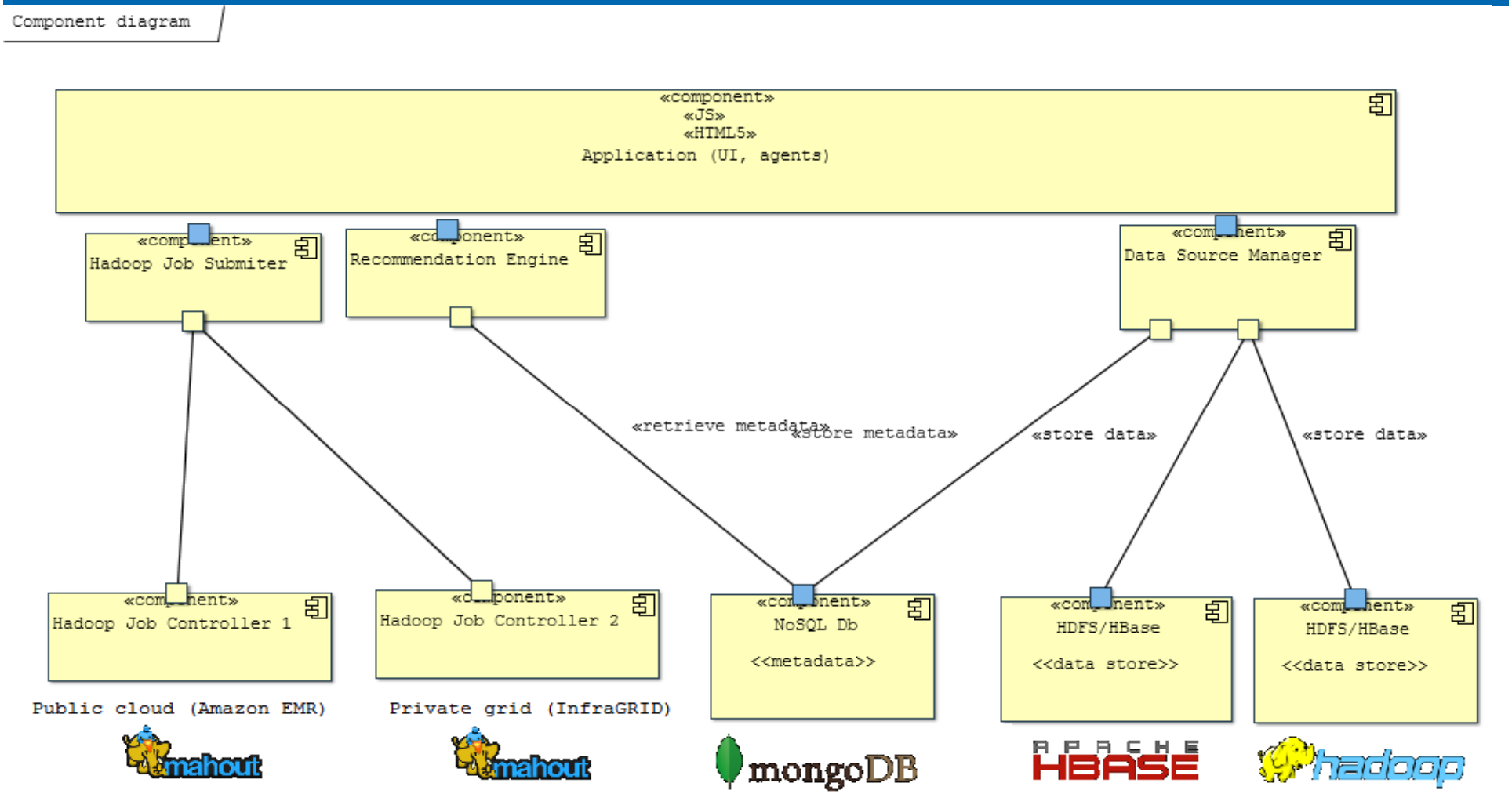
- Feedback missing => responsiveness
- Fine-tuning algorithms => customizable
- Simple solutions (easy to use) vs. Complex systems (expert skills)
- => usability
- Lack of recommendations => “smart”

Domain experts should be supported in data analytics tasks by user friendly, “smart”, customizable & responsive tools.

How?

- Distributed architecture for processing (Hadoop) and data (HBase / HDFS)
- Apache Mahout for effective parallel machine learning algorithms
- Lightweight web service (RESTful)
- NoSQL stores
- Recommendation engine based on ontologies for ML

How?


Component diagram
«component»
«JS»
«HTML5»
Application (UI, agents)
«component»
Hadoop Job Submiter
«component»
Recommendation Engine
«component»
Data Source Manager
«retrieve metadata»
«store metadata»
«store data»
«store data»
«component»
Hadoop Job Controller 1
«component»
Hadoop Job Controller 2
«component»
NoSQL Db
<<metadata>>
«component»
HDFS/HBase
<<data store>>
«component»
HDFS/HBase
<<data store>>
Public cloud (Amazon EMR)
Private grid (InfraGRID)
mahout
mahout
mongoDB
APACHE HBASE
hadoop

Want some?

- Enthusiast contributors are welcome
  - Architecture designers
  - Developers
  - Data providers

- Open-source project

Thank you