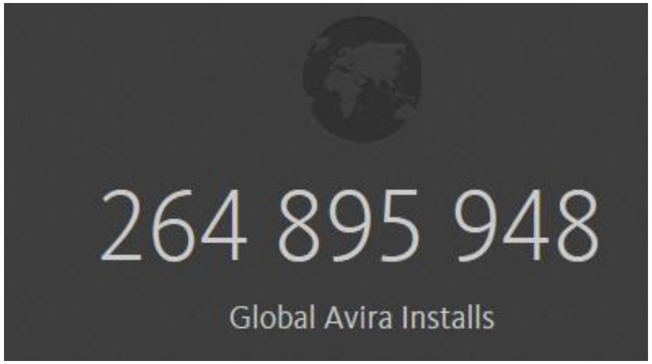
KNIME & Avira, or how I‘ve learned to love Big Data

Facts about Avira (AntiVir)

- 100 mio. customers
- „Extreme Reliability“
- 500 employees (Tettnang, San Francisco, Kuala Lumpur, Bucharest, Amsterdam)



264 895 948

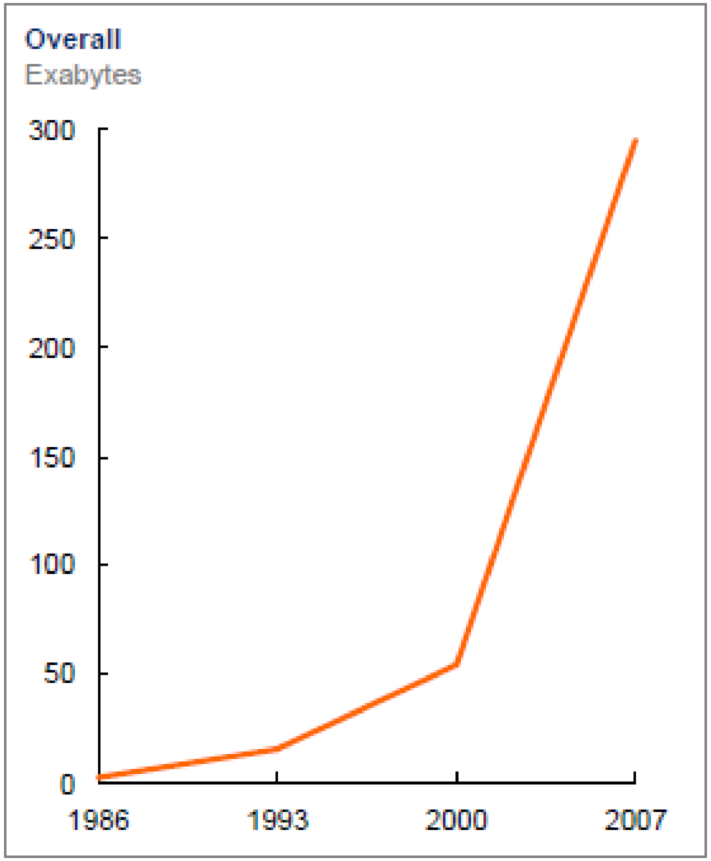
Global Avira Installs

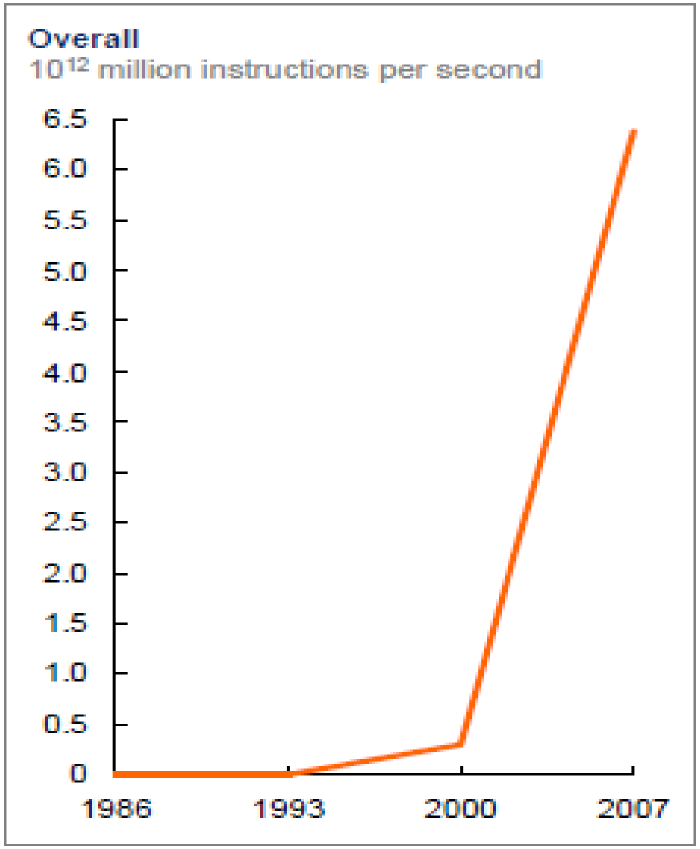
13.02.2014, 06:58

Company owner, Mr. Auerbach

Big Data: Why did Avira decided to invest?

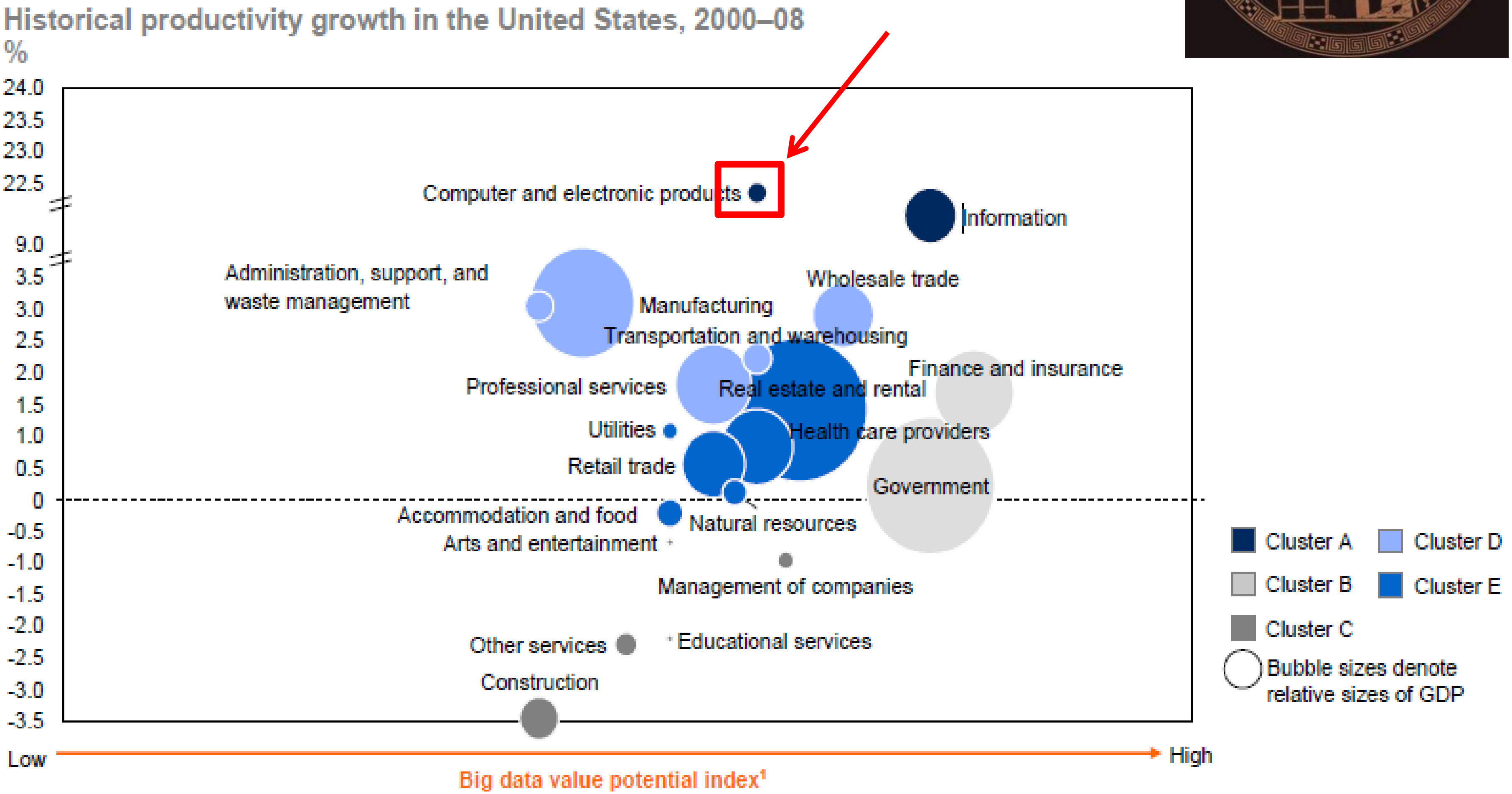
- Data storage has grown significantly after 2000
- Computation capacity has also risen sharply





SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Reason 1: Because the McKinsey prophets said it may bring us more money ☺



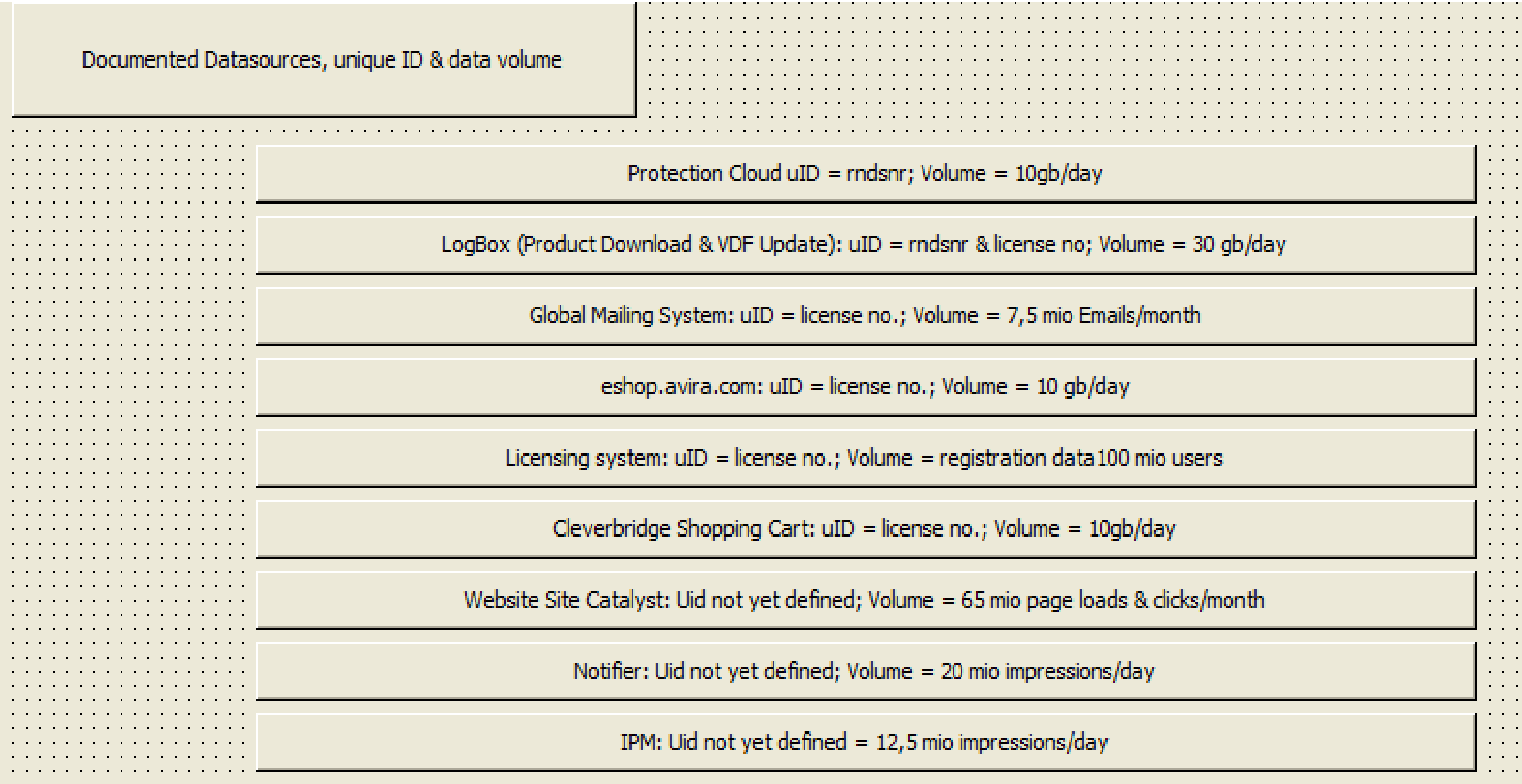
Historical productivity growth in the United States, 2000–08

%

SOURCE: US Bureau of Labor Statistics; McKinsey Global Institute analysis

Reason 2: Because we are sitting on the top of several dozens of terabytes of anonymized customer data.

The only challenge was to link this data and converge towards a single customer profile



Documented Datasources, unique ID & data volume

Protection Cloud uID = rndsnr; Volume = 10gb/day

LogBox (Product Download & VDF Update): uID = rndsnr & license no; Volume = 30 gb/day

Global Mailing System: uID = license no.; Volume = 7,5 mio Emails/month

eshop.avira.com: uID = license no.; Volume = 10 gb/day

Licensing system: uID = license no.; Volume = registration data100 mio users

Cleverbridge Shopping Cart: uID = license no.; Volume = 10gb/day

Website Site Catalyst: Uid not yet defined; Volume = 65 mio page loads & clicks/month

Notifier: Uid not yet defined; Volume = 20 mio impressions/day

IPM: Uid not yet defined = 12,5 mio impressions/day

Reason 3: Because we want to perform customer profilling & next best offer marketing to increase our margin

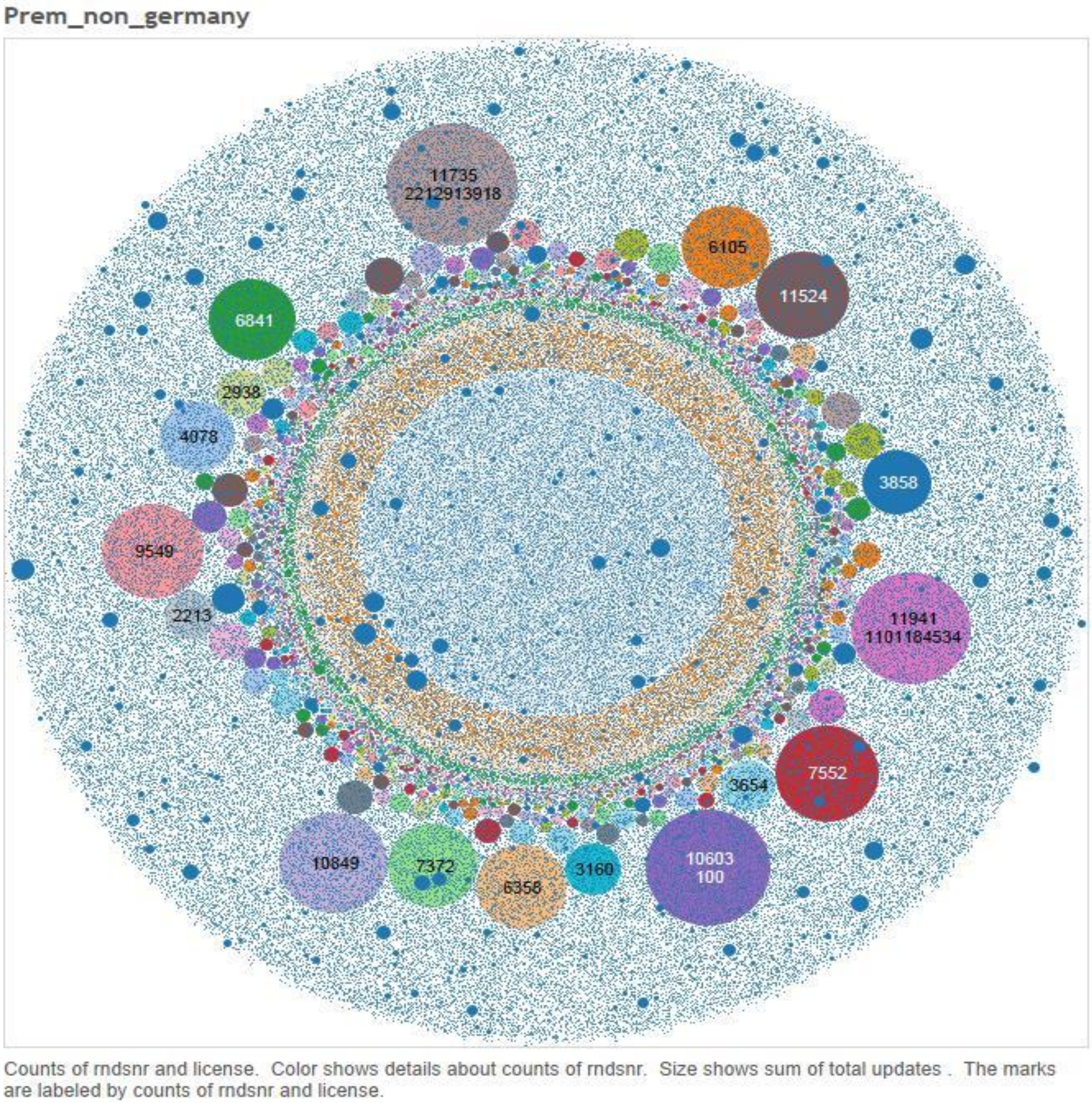
Business objectives

- Deliver the optimal price, features & messaging to each customer in order to maximize EBIT from sales of new and renewal licenses
- Protect long-term margins by making each offer timely and unrepeatable (e.g., unique to a specific customer, product, event)
- Learn the underlying mechanics of features and price-elasticity on the level of customer cells
- Enable business to optimize campaign portfolio (i.e., über-algorithm traffics campaigns)
- A/B testing in all customer touch-points

Technology requirements

- Ability to link a specific offer in a specific touch-point to specific customer.
- Machine-learning over all design dimensions to continually improve performance of the application
- Causal reporting to maximize the learning effect in the organization from algorithmic approaches to marketing and automation
- Create customer, product and behavioral tables from Avira's raw data within the dev. environment
- Create machine-learning algorithms optimizing the offer (price, features) per user-session
- Implement the services on development platform and place in listening mode to train
- Setup of the Hadoop framework (HDFS & MapReduce) & Couchbase, KNIME & Impala

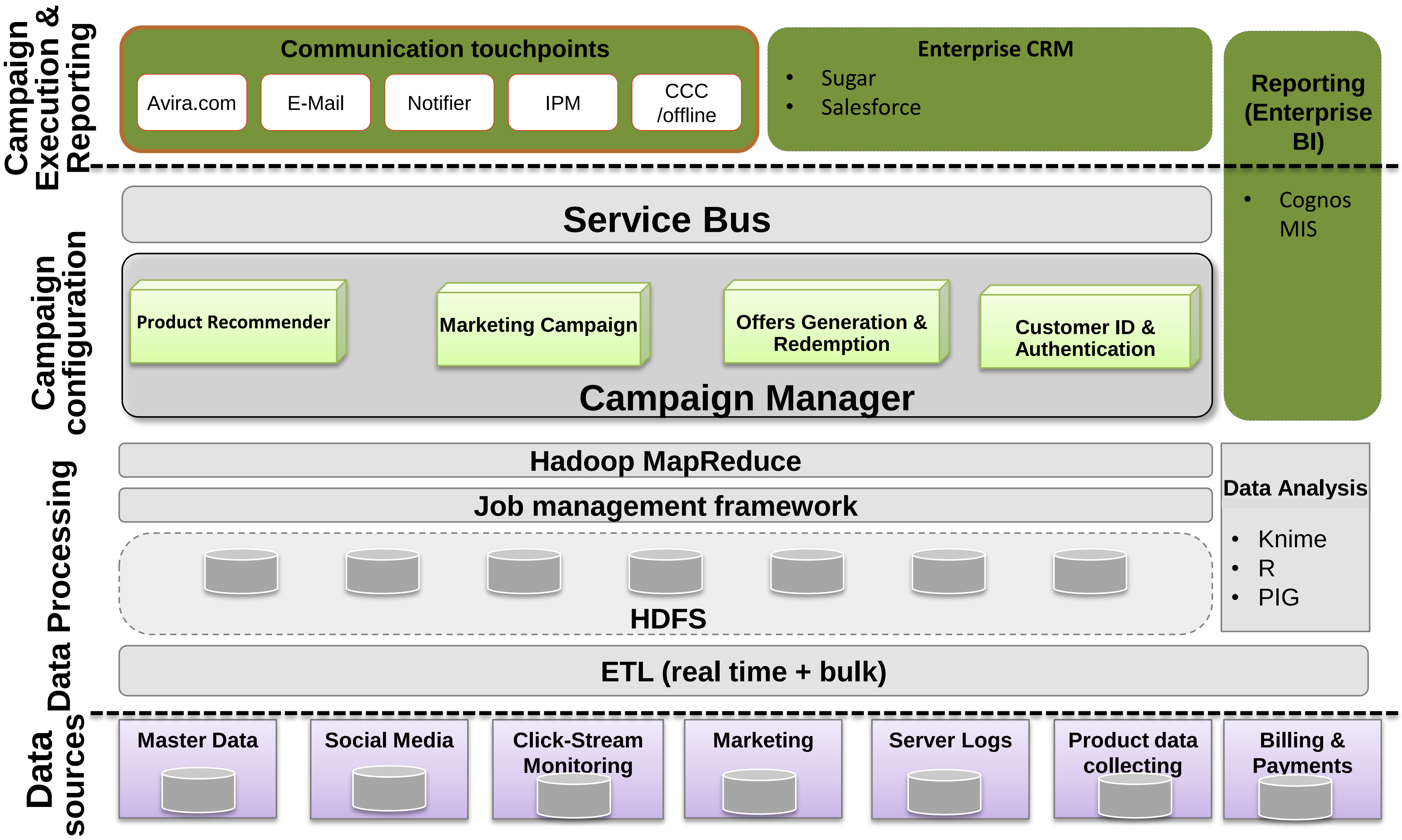
Reason 4: Because it‘s fun (isn‘t this beautiful?)



Prem_non_germany

Counts of rndsnr and license. Color shows details about counts of rndsnr. Size shows sum of total updates . The marks are labeled by counts of rndsnr and license.

Our business/architectural vision



Campaign Execution & Reporting

Communication touchpoints

Avira.com

E-Mail

Notifier

IPM

CCC /offline

Enterprise CRM

- Sugar
- Salesforce

Reporting (Enterprise BI)

- Cognos MIS

Campaign configuration

Service Bus

Product Recommender

Marketing Campaign

Offers Generation & Redemption

Customer ID & Authentication

Campaign Manager

Data Processing

Hadoop MapReduce

Job management framework

HDFS

Data Analysis

- Knime
- R
- PIG

ETL (real time + bulk)

Data sources

Master Data

Social Media

Click-Stream Monitoring

Marketing

Server Logs

Product data collecting

Billing & Payments

Our Daily Data

- Website logs: 5.000.000 lines
- Installation logs: 2.200.000 lines
- InProduct Messaging: 43.000.000 lines
- Download/Updater logs: 2.000.000.000 lines

60GB of daily compressed data

or around 2TB of monthly data

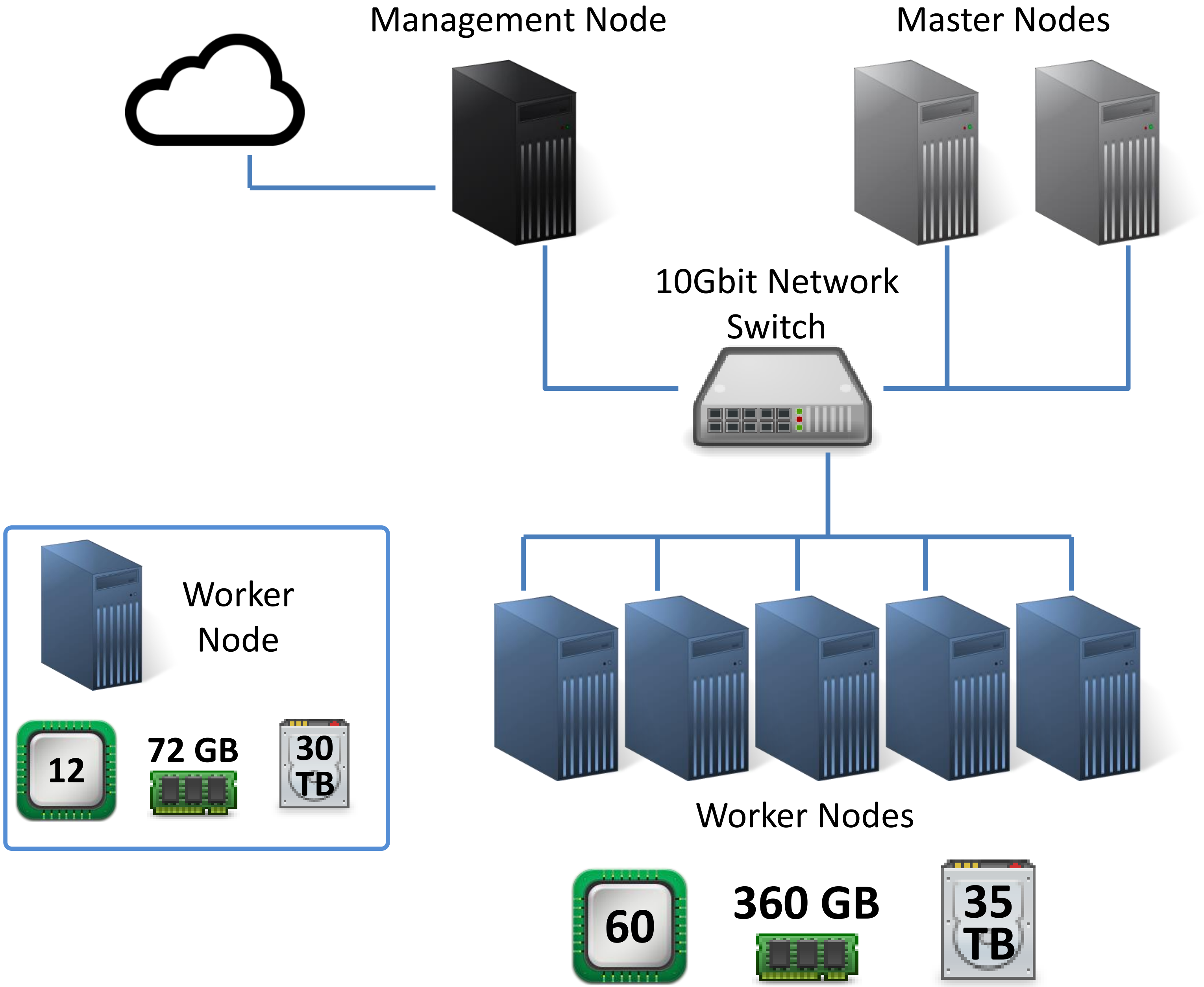
Finding the Right Tool

- Our data is:
  - Unstructured; Messy
  - Coming from all kinds of sources:

    log files, log tables, relational databases

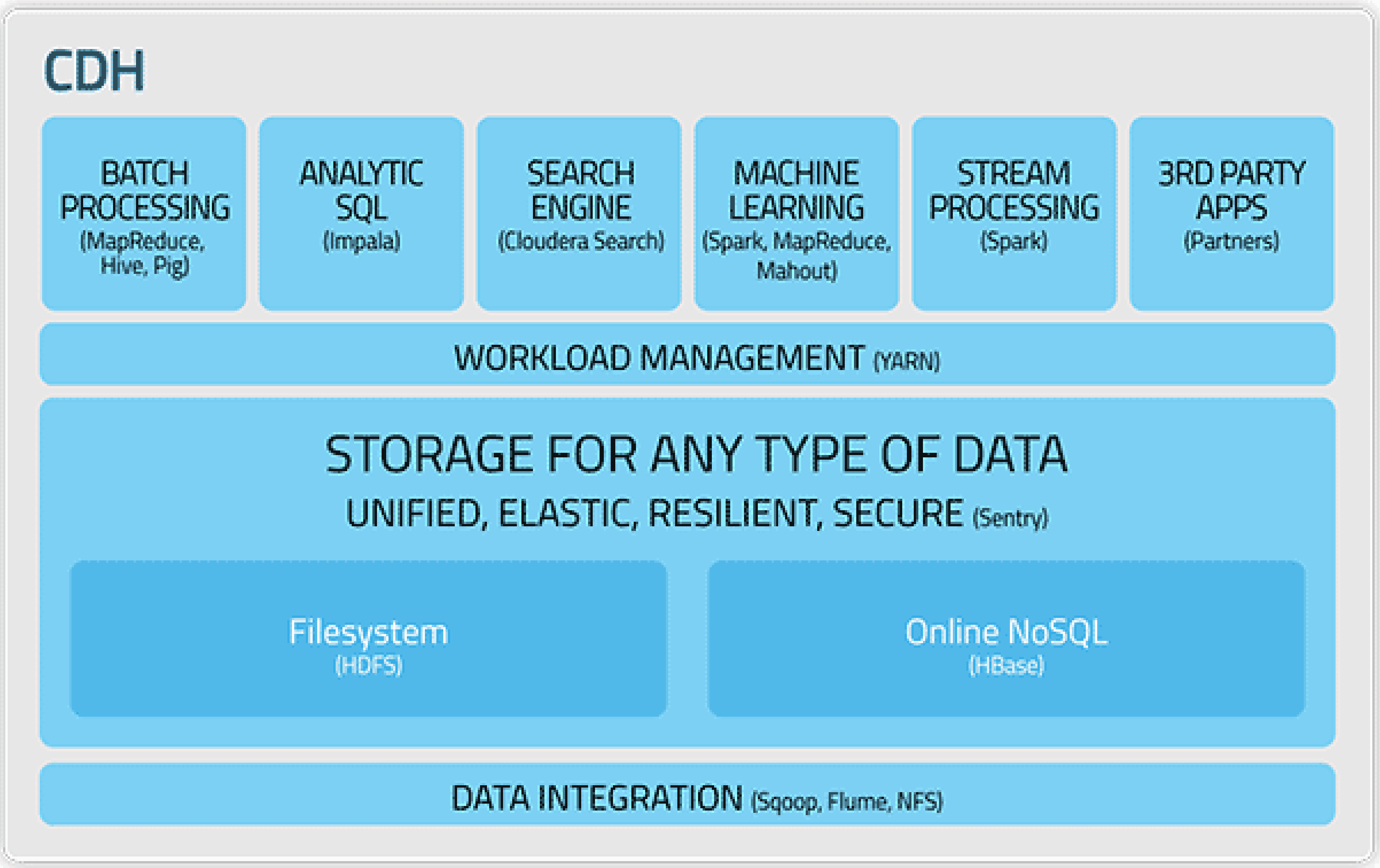
- We want to:
  - Gather and store historical data
  - Process huge amounts of it
  - Support both batch and real-time operations

hadoop


Management Node
Master Nodes
10Gbit Network Switch
Worker Node
12
72 GB
30 TB
Worker Nodes
60
360 GB
35 TB

cloudera®

Ask Bigger Questions



CDH

BATCH PROCESSING (MapReduce, Hive, Pig)

ANALYTIC SQL (Impala)

SEARCH ENGINE (Cloudera Search)

MACHINE LEARNING (Spark, MapReduce, Mahout)

STREAM PROCESSING (Spark)

3RD PARTY APPS (Partners)

WORKLOAD MANAGEMENT (YARN)

STORAGE FOR ANY TYPE OF DATA

UNIFIED, ELASTIC, RESILIENT, SECURE (Sentry)

Filesystem (HDFS)

Online NoSQL (HBase)

DATA INTEGRATION (Sqoop, Flume, NFS)

Using The Right Tool

MapReduce

- Full control over how the data is processed
- Works on structured and unstructured data
- Good for very complex business logic

- You have to write Java code
- Restricted to the MapReduce programming model
- Some things are difficult to code (JOINs, custom sorting)

Hive

- You write SQL-like queries
- Great for ad-hoc queries, data exploration
- Very fast development

- Works only on structured data
- Gets ugly if the business logic is complex

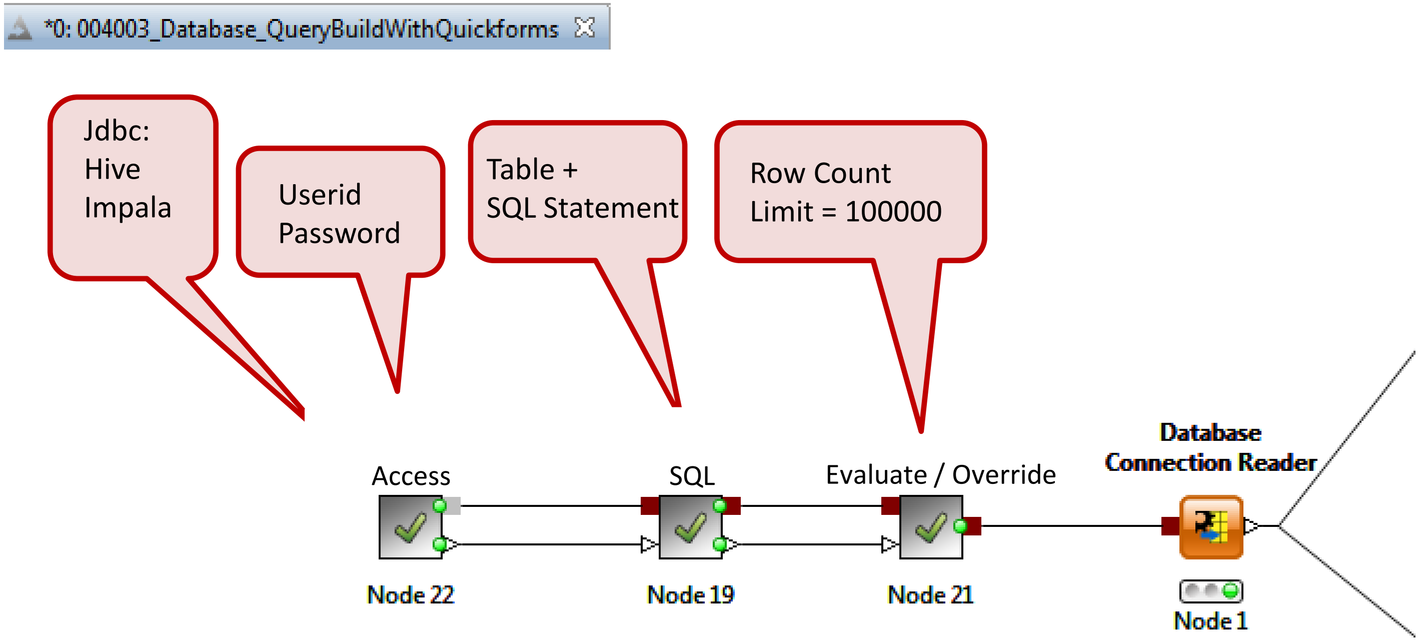
Impala

cloudera®
IMPALA

- Like Hive but A LOT faster
- Runs directly in memory
- Delivers almost real-time results

- Limited to in-memory processing
- Unreliable

KNIME and Hadoop at Avira


*0: 004003_Database_QueryBuildWithQuickforms
Access
SQL
Evaluate / Override
Database Connection Reader
Node 22
Node 19
Node 21
Node 1

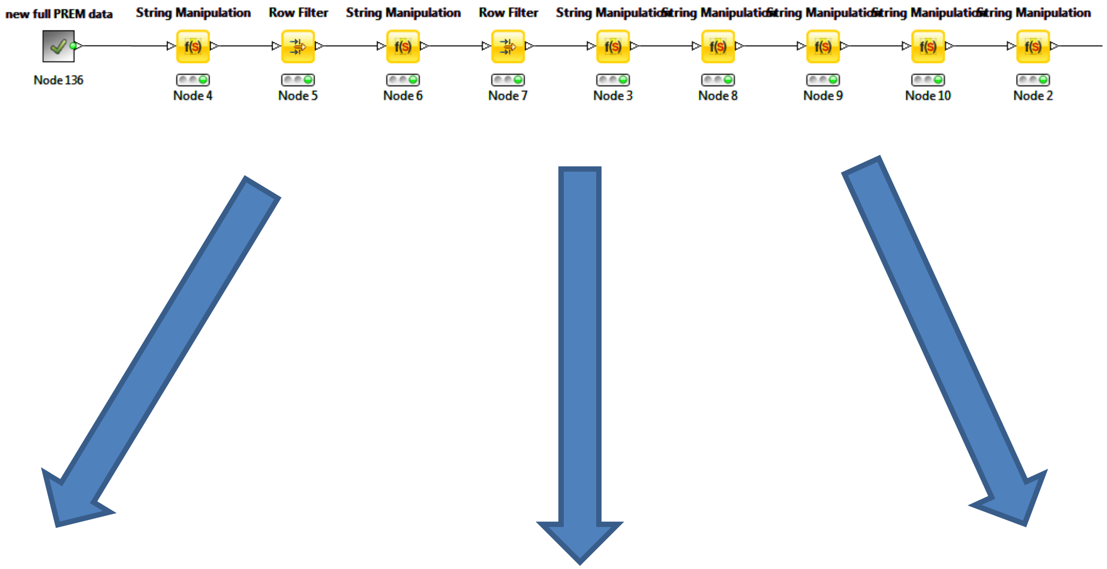
Jdbc:
Hive
Impala

Userid
Password

Table +
SQL Statement

Row Count
Limit = 100000

Knime for Understanding the Data



MapReduce

Hive / Impala

KNIME

Some Big Data Practicalities

Use Crisp-DM !

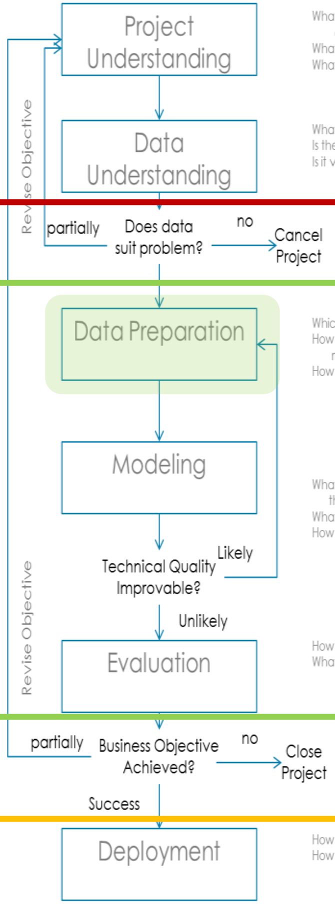
Big Data Methods
Never Needed

Decision: Would
Big Data methods add value?

Big Data Methods
Possibly useful

Decision: Would
Big Data methods add value?

Big Data Methods
Possibly useful



What exactly is the problem, the expected benefit?
What should a solution look like?
What is known about the domain

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How well does the model perform technically?

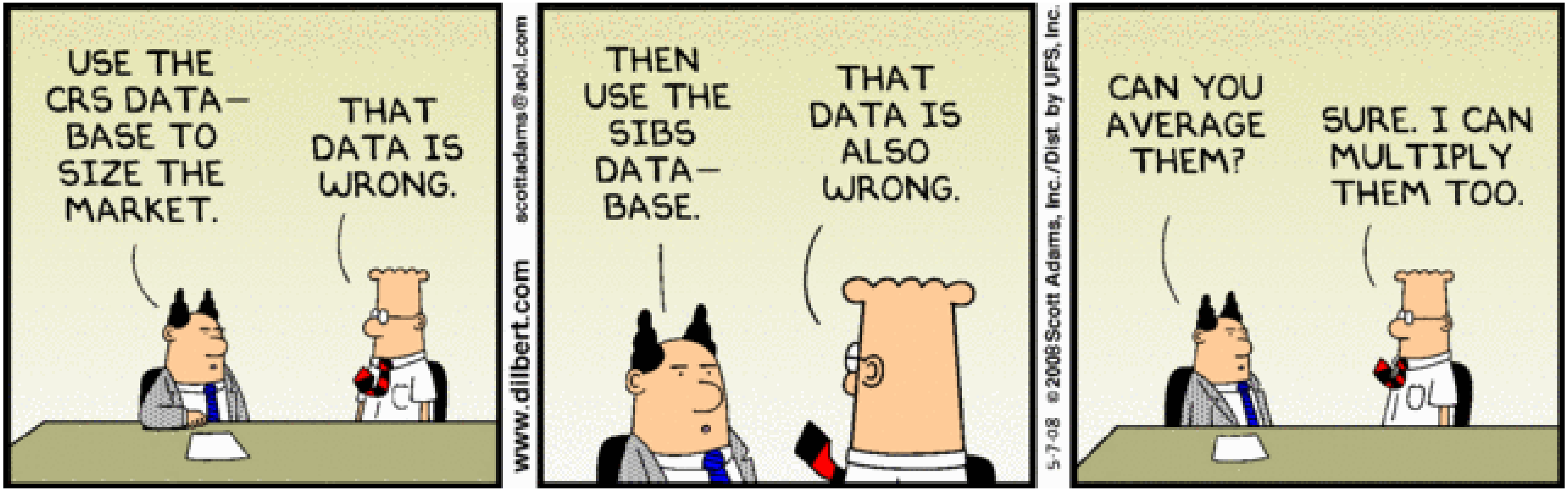
How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

Guide to Intelligent Data Analysis, Berthold, 2010

Why did we decided to go with KNIME?

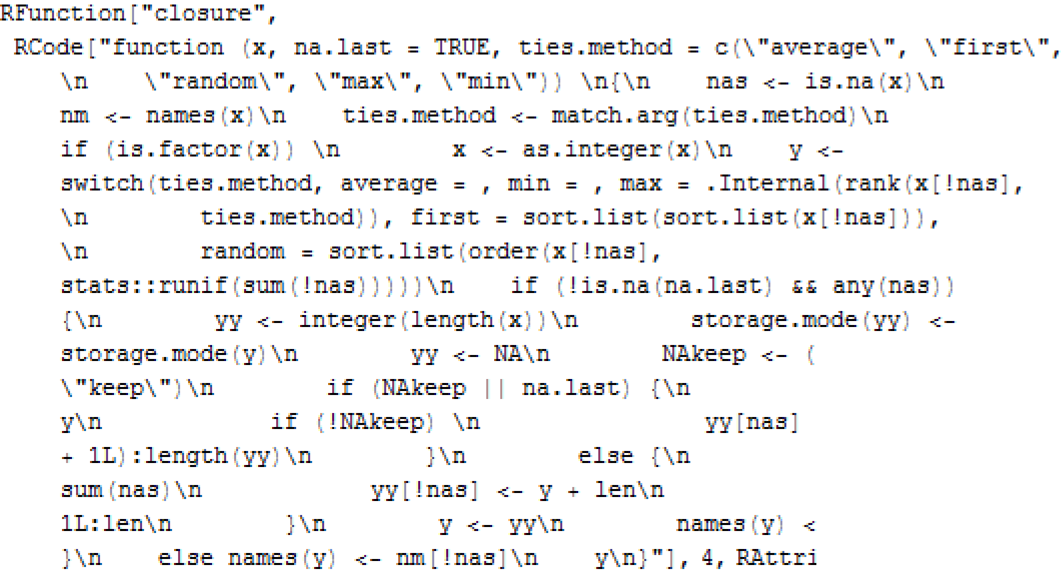
- The dark side of the moon: a typical symptom for home-grown business applications.


USE THE CRS DATA-BASE TO SIZE THE MARKET.
THAT DATA IS WRONG.
www.dilbert.com scottadams@aol.com
THEN USE THE SIBS DATA-BASE.
THAT DATA IS ALSO WRONG.
5-7-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.
CAN YOU AVERAGE THEM?
SURE. I CAN MULTIPLY THEM TOO.

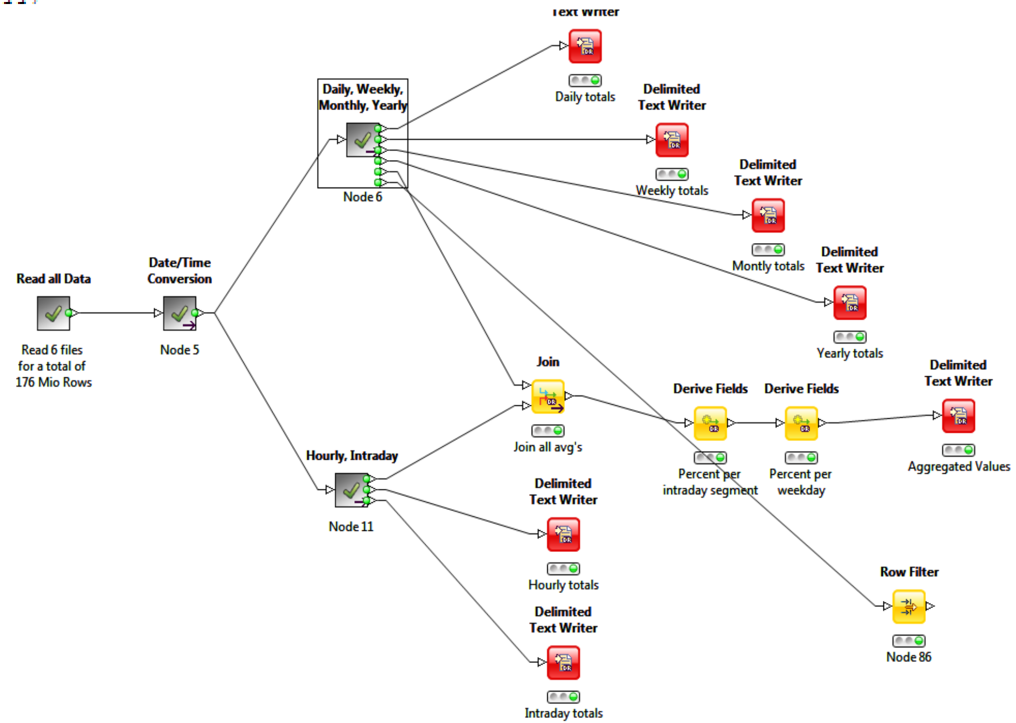
KNIME helped us to tie a „knot“ for the multiple uncorrelated data points and create customer 360 tables

Why did we decided to go with KNIME?

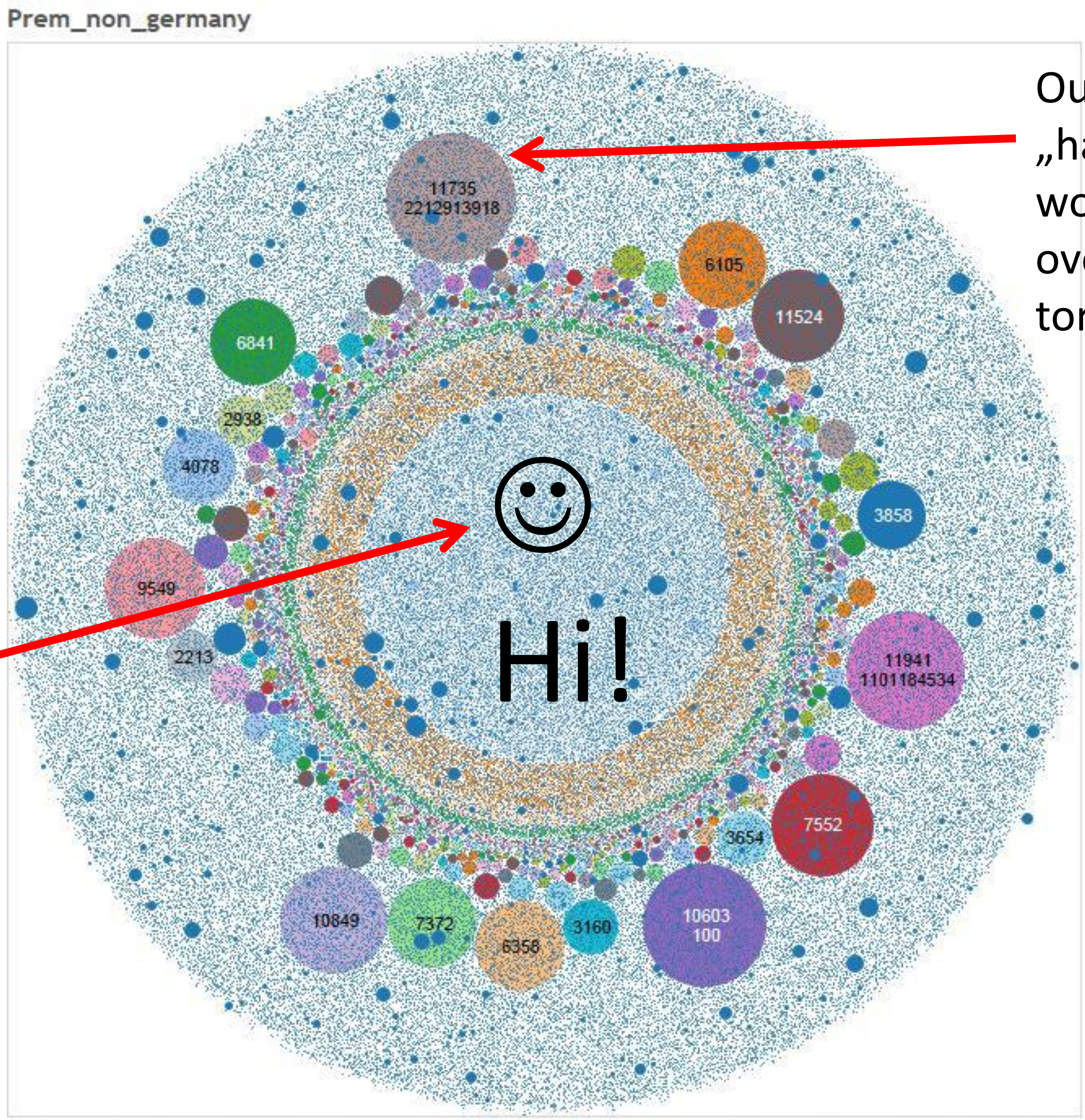
It helped us move from code based data mining towards workflow based analytics; Analytics for everyone, easy to explain to all management/company levels



```
RFunction["closure",
 RCode["function (x, na.last = TRUE, ties.method = c(\"average\", \"first\",
    \n    \"random\", \"max\", \"min\")) \n{\n    nas <- is.na(x)\n
    nm <- names(x)\n    ties.method <- match.arg(ties.method)\n
    if (is.factor(x)) \n        x <- as.integer(x)\n    y <-
    switch(ties.method, average = , min = , max = .Internal(rank(x[!nas],
    \n        ties.method)), first = sort.list(sort.list(x[!nas])),
    \n        random = sort.list(order(x[!nas],
    stats::runif(sum(!nas)))))\n    if (!is.na(na.last) && any(nas))
    {\n        yy <- integer(length(x))\n        storage.mode(yy) <-
    storage.mode(y)\n        yy <- NA\n        NAkeep <- (
    \"keep\")\n        if (NAkeep || na.last) {\n
    y\n            if (!NAkeep) \n                yy[nas]
    + 1L):length(yy)\n        }\n        else {\n
    sum(nas)\n            yy[!nas] <- y + len\n
    1L:len\n        }\n        y <- yy\n        names(y) <
    }\n    else names(y) <- nm[!nas]\n    y\n}"], 4, RAttri
```



Some results: using KNIME & Tableau we‘ve managed to perform forensycs and license outlier analysis



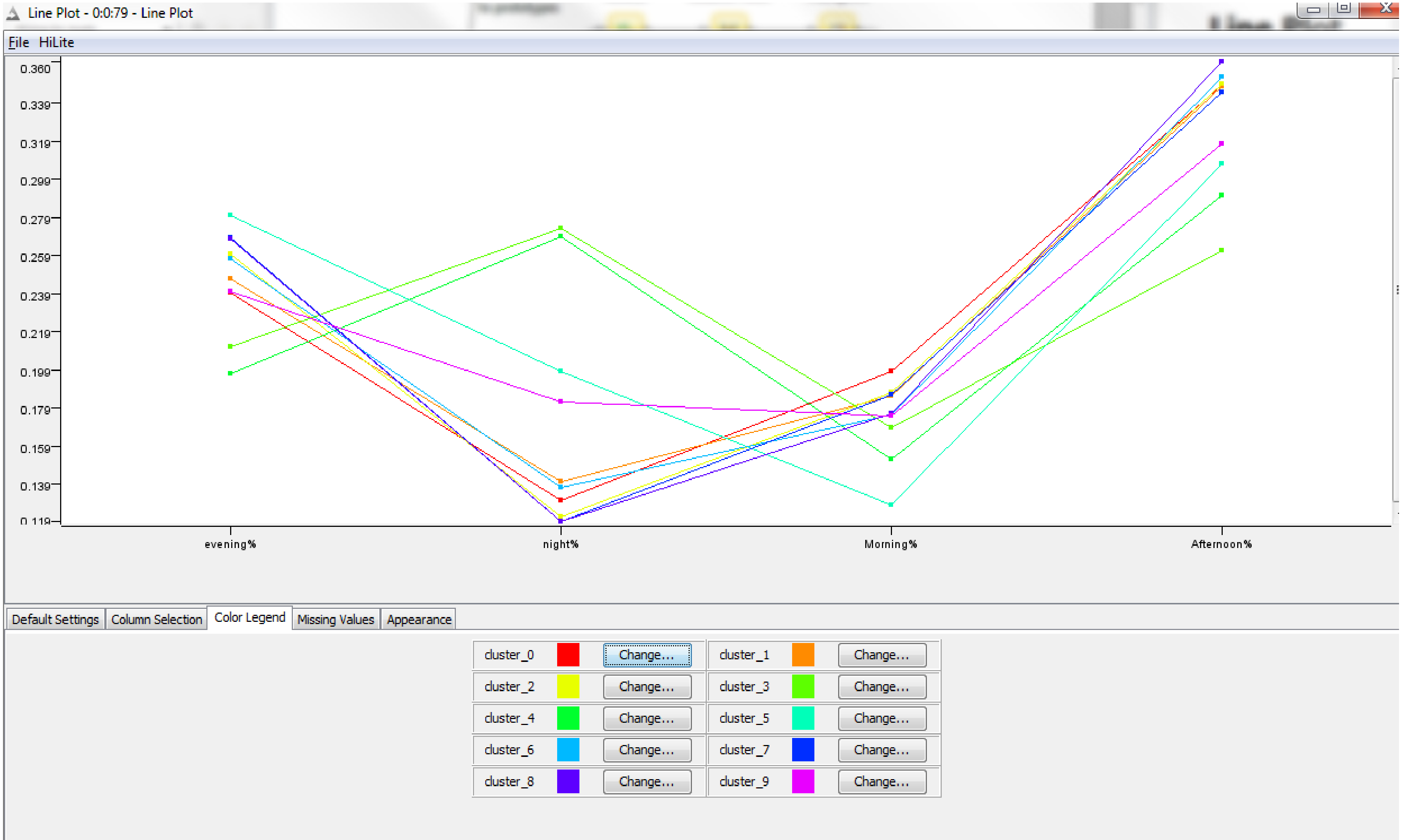
Counts of rndsnr and license. Color shows details about counts of rndsnr. Size shows sum of total updates . The marks are labeled by counts of rndsnr and license.

Our notorious „hacker friend“ working for the overall good of torrent visitors

This is me using Avira in Sep 2013

Running k-means in KNIME to identify relevant clusters for Germany by looking at their antivirus software update behaviour

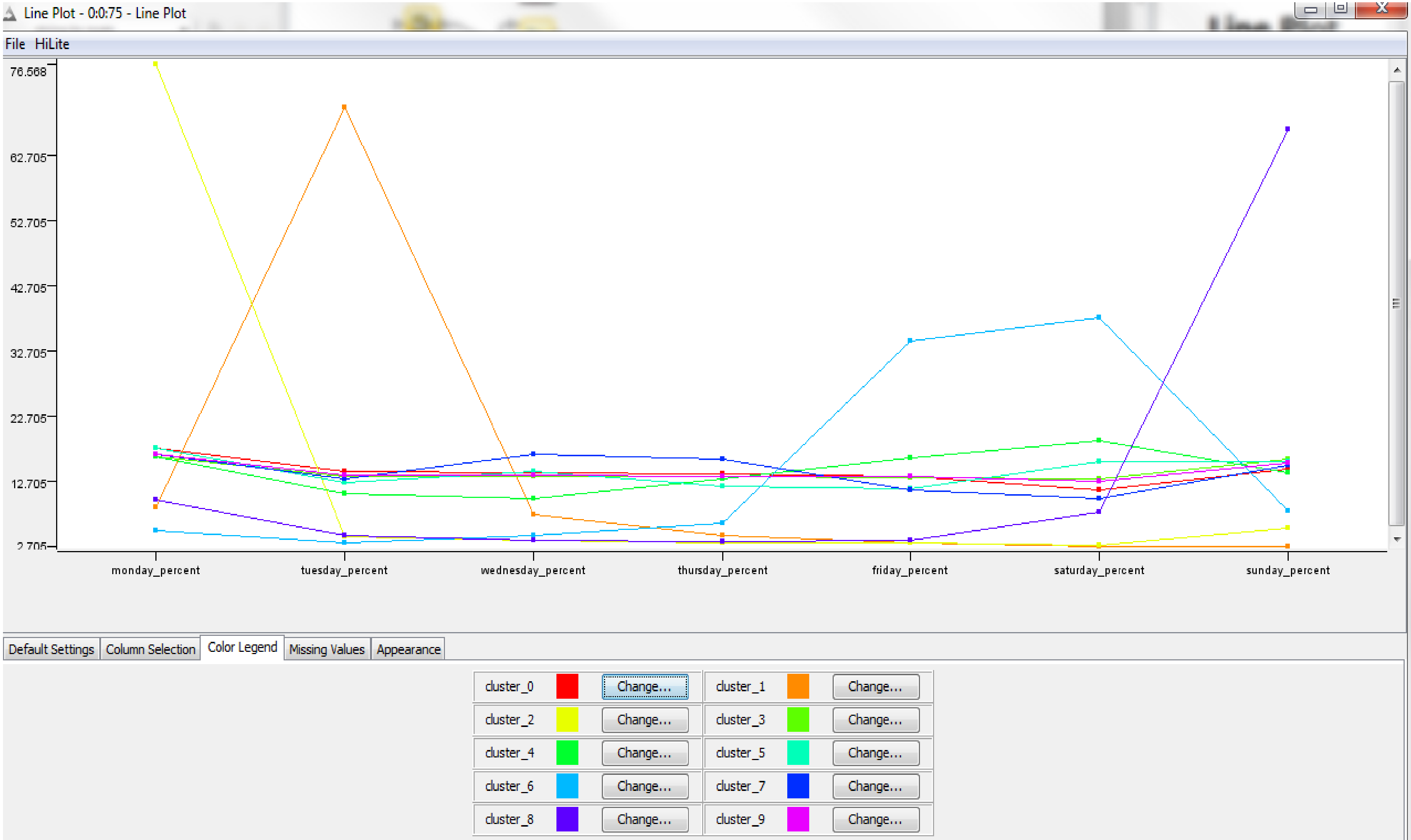
Graph of clusters by time of day % usage:



cluster_0	🟥	Change...	cluster_1	🟧	Change...
cluster_2	🟨	Change...	cluster_3	🟩	Change...
cluster_4	🟩	Change...	cluster_5	🟦	Change...
cluster_6	🟦	Change...	cluster_7	🟦	Change...
cluster_8	🟪	Change...	cluster_9	🟪	Change...

Running k-means in KNIME to identify relevant clusters for Germany by looking at their antivirus software update behaviour

Graph of clusters by daily % usage

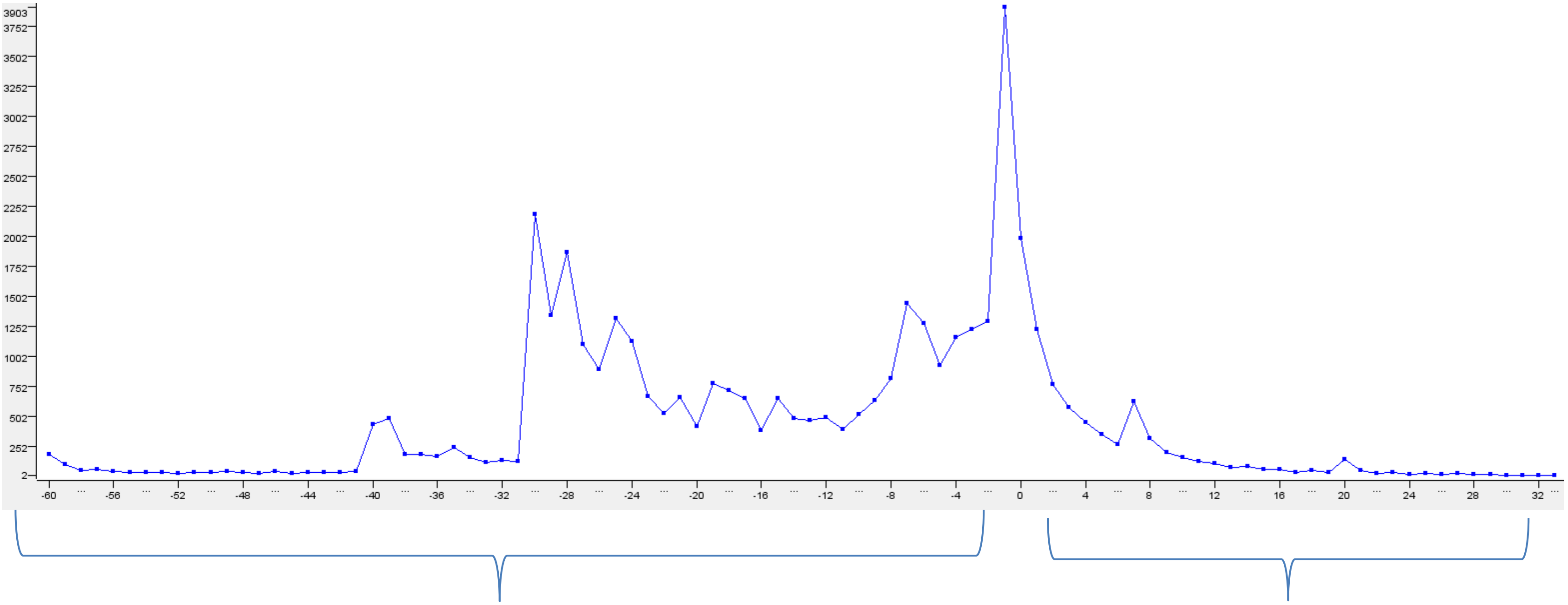


cluster_0	🟥	Change...	cluster_1	🟧	Change...
cluster_2	🟨	Change...	cluster_3	🟩	Change...
cluster_4	🟩	Change...	cluster_5	🟩	Change...
cluster_6	🟦	Change...	cluster_7	🟦	Change...
cluster_8	🟪	Change...	cluster_9	🟪	Change...

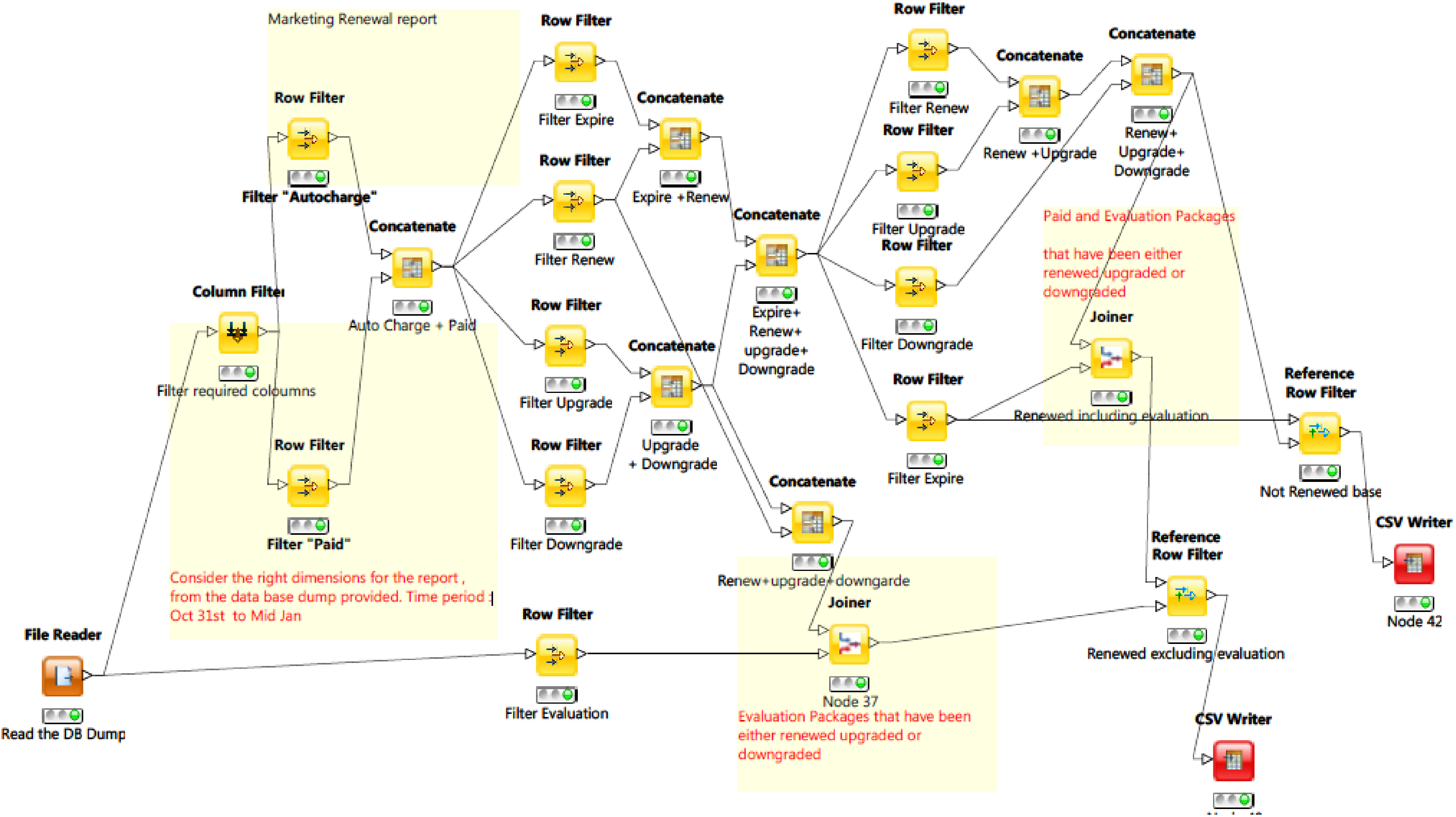
Using KNIME to identify the real License Renewal pattern of our customers

The timing of renewing an Avira license in %

“Reason 6” = 12%



Using KNIME to do standardized reporting of our license renewal metrics



Marketing Renewal report

Consider the right dimensions for the report , from the data base dump provided. Time period : Oct 31st to Mid Jan

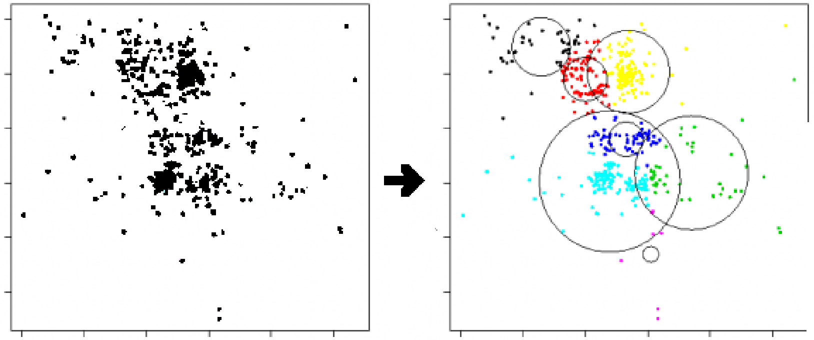
Paid and Evaluation Packages

that have been either renewed,upgraded or downgraded

Evaluation Packages that have been either renewed upgraded or downgraded

Next steps

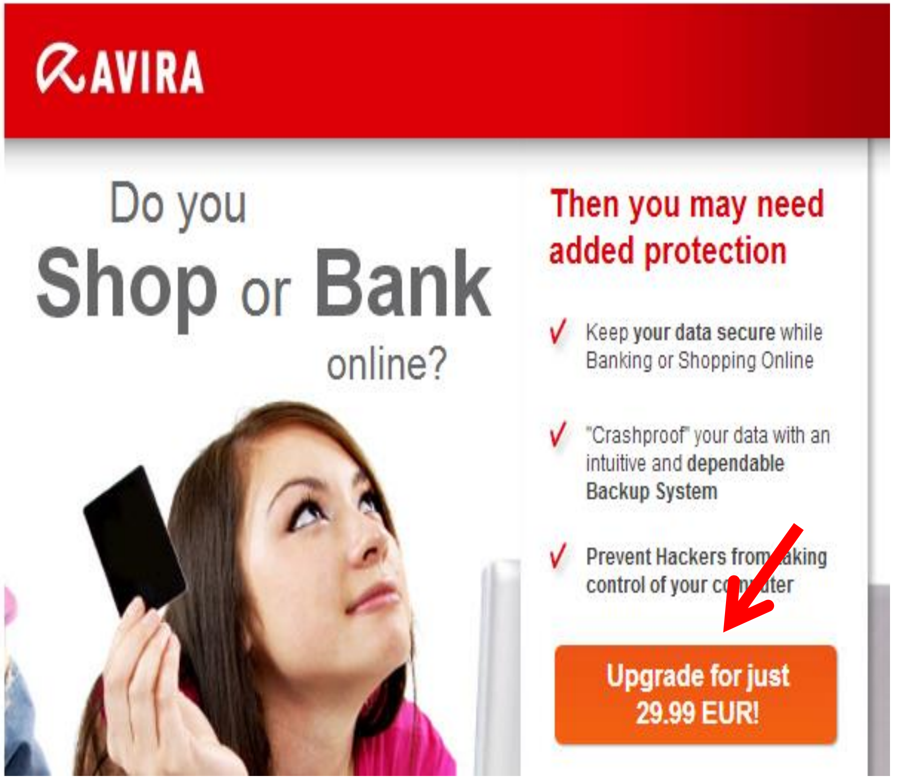
- Identify unknown groups of customers by allowing the machine to find patterns in data for creating special association rules/product recommendations & next best offer; test & train in KNIME, real-time model execution in Couchbase;



Raw Data

Clustered Data

NBO via Email



AVIRA

Do you Shop or Bank online?

Then you may need added protection

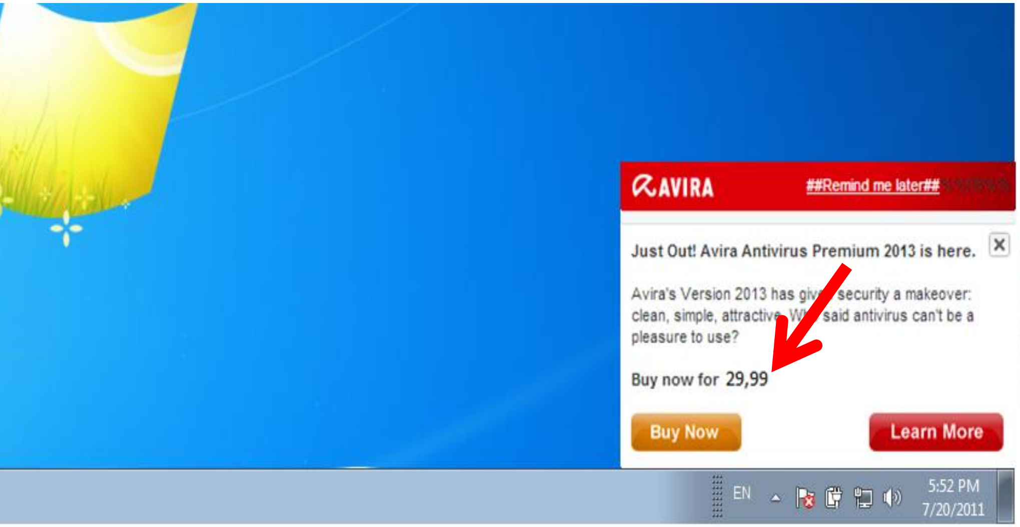
✓ Keep your data secure while Banking or Shopping Online

✓ "Crashproof" your data with an intuitive and dependable Backup System

✓ Prevent Hackers from aking control of your co uter

Upgrade for just 29.99 EUR!

NBO via In App Messaging



AVIRA ##Remind me later##

Just Out! Avira Antivirus Premium 2013 is here.

Avira's Version 2013 has giv security a makeover: clean, simple, attractive. W said antivirus can't be a pleasure to use?

Buy now for 29,99

Thank you.

radu.pastia@avira.com/ phil.winters@knime.com/ florin.veringa@avira.com