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)

Company owner, Mr. Auerbach
Big Data: Why did Avira decide 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 😊
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.

<table>
<thead>
<tr>
<th>Documented Datasources, unique ID &amp; data volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protection Cloud uID = rndsnr; Volume = 10gb/day</td>
</tr>
<tr>
<td>LogBox (Product Download &amp; VDF Update): uID = rndsnr &amp; license no; Volume = 30 gb/day</td>
</tr>
<tr>
<td>Global Mailing System: uID = license no.; Volume = 7,5 mio Emails/month</td>
</tr>
<tr>
<td>eshop.avira.com: uID = license no.; Volume = 10 gb/day</td>
</tr>
<tr>
<td>Licensing system: uID = license no.; Volume = registration data 100 mio users</td>
</tr>
<tr>
<td>Cleverbridge Shopping Cart: uID = license no.; Volume = 10gb/day</td>
</tr>
<tr>
<td>Website Site Catalyst: Uid not yet defined; Volume = 65 mio page loads &amp; clicks/month</td>
</tr>
<tr>
<td>Notifier: Uid not yet defined; Volume = 20 mio impressions/day</td>
</tr>
<tr>
<td>IPM: Uid not yet defined = 12,5 mio impressions/day</td>
</tr>
</tbody>
</table>
Reason 3: Because we want to perform customer profiling & next best offer marketing to increase our margin

<table>
<thead>
<tr>
<th>Business objectives</th>
<th>Technology requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Deliver the optimal price, features &amp; messaging to each customer in order to maximize EBIT from sales of new and renewal licenses</td>
<td></td>
</tr>
<tr>
<td>• Protect long-term margins by making each offer timely and unrepeatable (e.g., unique to a specific customer, product, event)</td>
<td></td>
</tr>
<tr>
<td>• Learn the underlying mechanics of features and price-elasticity on the level of customer cells</td>
<td></td>
</tr>
<tr>
<td>• Enable business to optimize campaign portfolio (i.e., über-algorithm traffics campaigns)</td>
<td></td>
</tr>
<tr>
<td>• A/B testing in all customer touch-points</td>
<td></td>
</tr>
<tr>
<td>• Ability to link a specific offer in a specific touch-point to specific customer.</td>
<td></td>
</tr>
<tr>
<td>• Machine-learning over all design dimensions to continually improve performance of the application</td>
<td></td>
</tr>
<tr>
<td>• Causal reporting to maximize the learning effect in the organization from algorithmic approaches to marketing and automation</td>
<td></td>
</tr>
<tr>
<td>• Create customer, product and behavioral tables from Avira’s raw data within the dev. environment</td>
<td></td>
</tr>
<tr>
<td>• Create machine-learning algorithms optimizing the offer (price, features) per user-session</td>
<td></td>
</tr>
<tr>
<td>• Implement the services on development platform and place in listening mode to train</td>
<td></td>
</tr>
<tr>
<td>• Setup of the Hadoop framework (HDFS &amp; MapReduce) &amp; Couchbase, KNIME &amp; Impala</td>
<td></td>
</tr>
</tbody>
</table>
Reason 4: Because it’s fun (isn’t this beautiful?)
Our business/architectural vision

- **Communication touchpoints**
  - Avira.com
  - E-Mail
  - Notifier
  - IPM
  - CCC (offline)

- **Enterprise CRM**
  - Sugar
  - Salesforce

- **Reporting (Enterprise BI)**
  - Cognos MIS

- **Service Bus**
  - Product Recommender
  - Marketing Campaign
  - Offers Generation & Redemption
  - Customer ID & Authentication

- **Campaign Manager**

- **Hadoop MapReduce**

- **Job management framework**

- **ETL (real time + bulk)**

- **Data Analysis**
  - Knime
  - R
  - PIG

- **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
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

• 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

- Jdbc: Hive Impala
- Userid Password
- Table + SQL Statement
- Row Count Limit = 100000

Diagram:

- Node 22: Access
- Node 19: SQL
- Node 21: Evaluate / Override
- Database Connection Reader

Diagram showing data flow and connections between nodes.
Knime for Understanding the Data

MapReduce  
Hive / Impala  
KNIME
Some Big Data Practicalities

Use Crisp-DM!

Big Data Methods
Never Needed

Big Data Methods
Possibly useful

Big Data Methods
Possibly useful

Decision: Would Big Data methods add value?
Why did we decided to go with KNIME?

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

KNIME helped us to tie a „knot“ for the multiple uncorrelated data points and create customer 360 tables.
Why did we decide 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.
Some results: using KNIME & Tableau we’ve managed to perform forensyscs and license outlier analysis.

This is me using Avira in Sep 2013.

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

Consider the right dimensions for the report, from the database dump provided. Time period Oct 31st to Mid Jan.
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](image1)

![Clustered Data](image2)

NBO via Email

Do you Shop or Bank online?

- Keep your data secure while Banking or Shopping Online
- "Crashproof" your data with an intuitive and dependable Backup System
- Prevent Hackers from taking control of your computer

Upgrade for just 29.99 EUR!
Thank you.

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