TNG TECHNOLOGY CONSULTING

BIG DATA - GET MORE OUT OF YOUR DATA

Christoph Stock @christophstock
Bernhard Bock @bbock23

Big Data is...

data, which

- is so huge or
- changes so fast or
- has no well definied structure

so that it cannot be processed
with classical business intelligence tools.

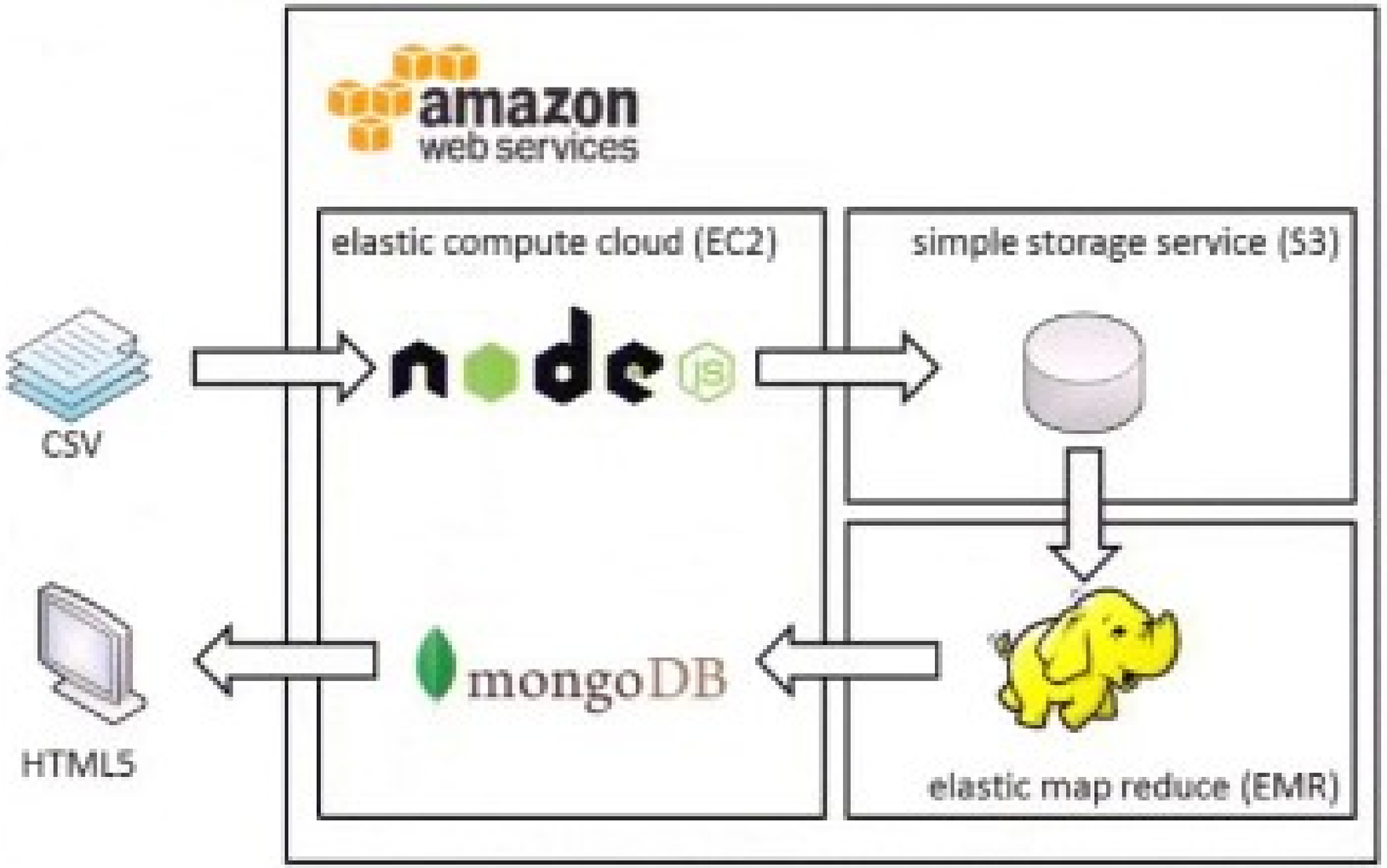
Since this talk has started

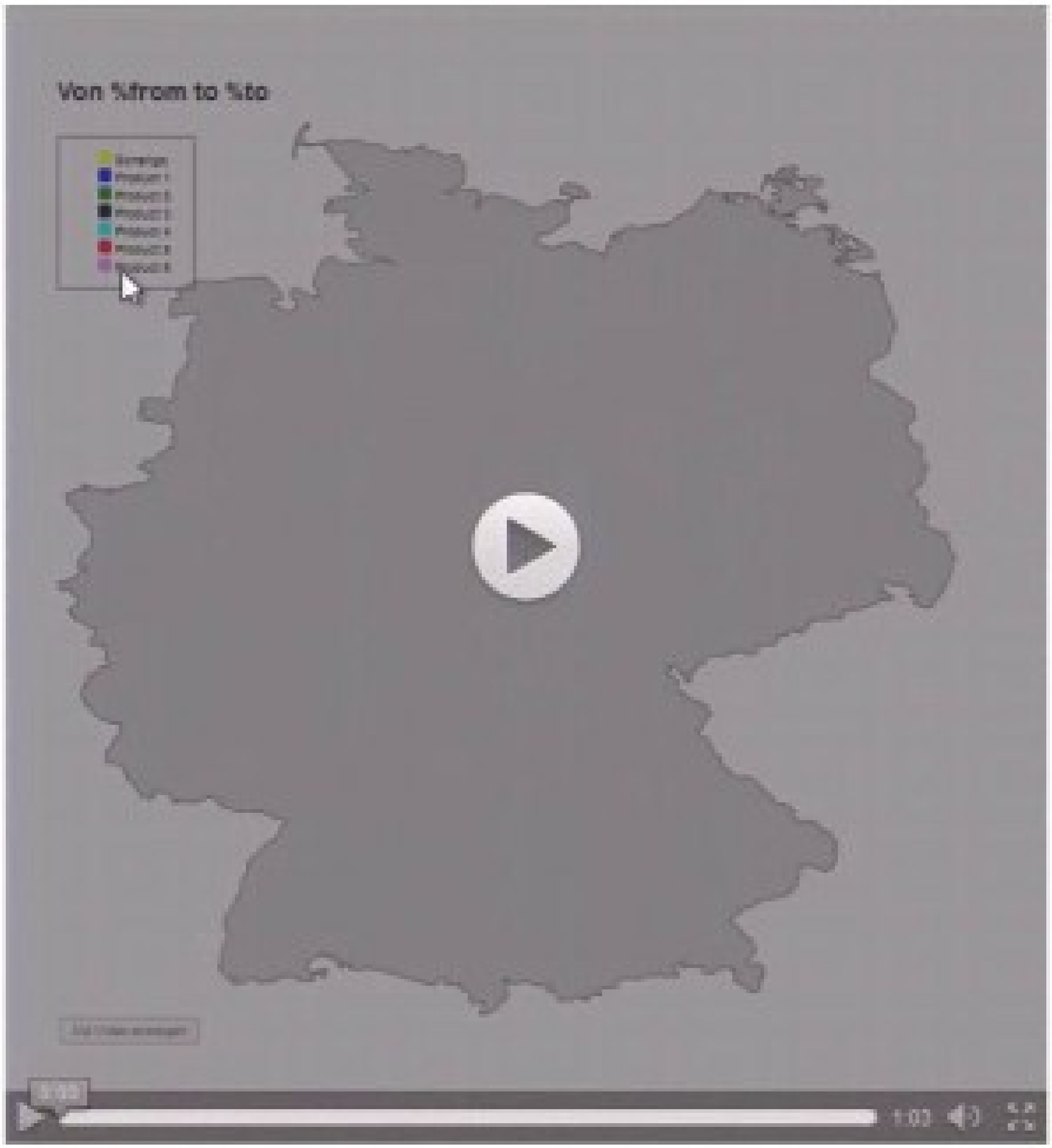
- 900000 photos were uploaded to Facebook
- 3000 new Android smart phones were activited

Usecase 1: Realtime Marketing

- Marketing has tools to steer on short timescales
- Datawarehouse provides only historic reports
- Idea: Use existing event stream and display on a map

Application flow (POC)

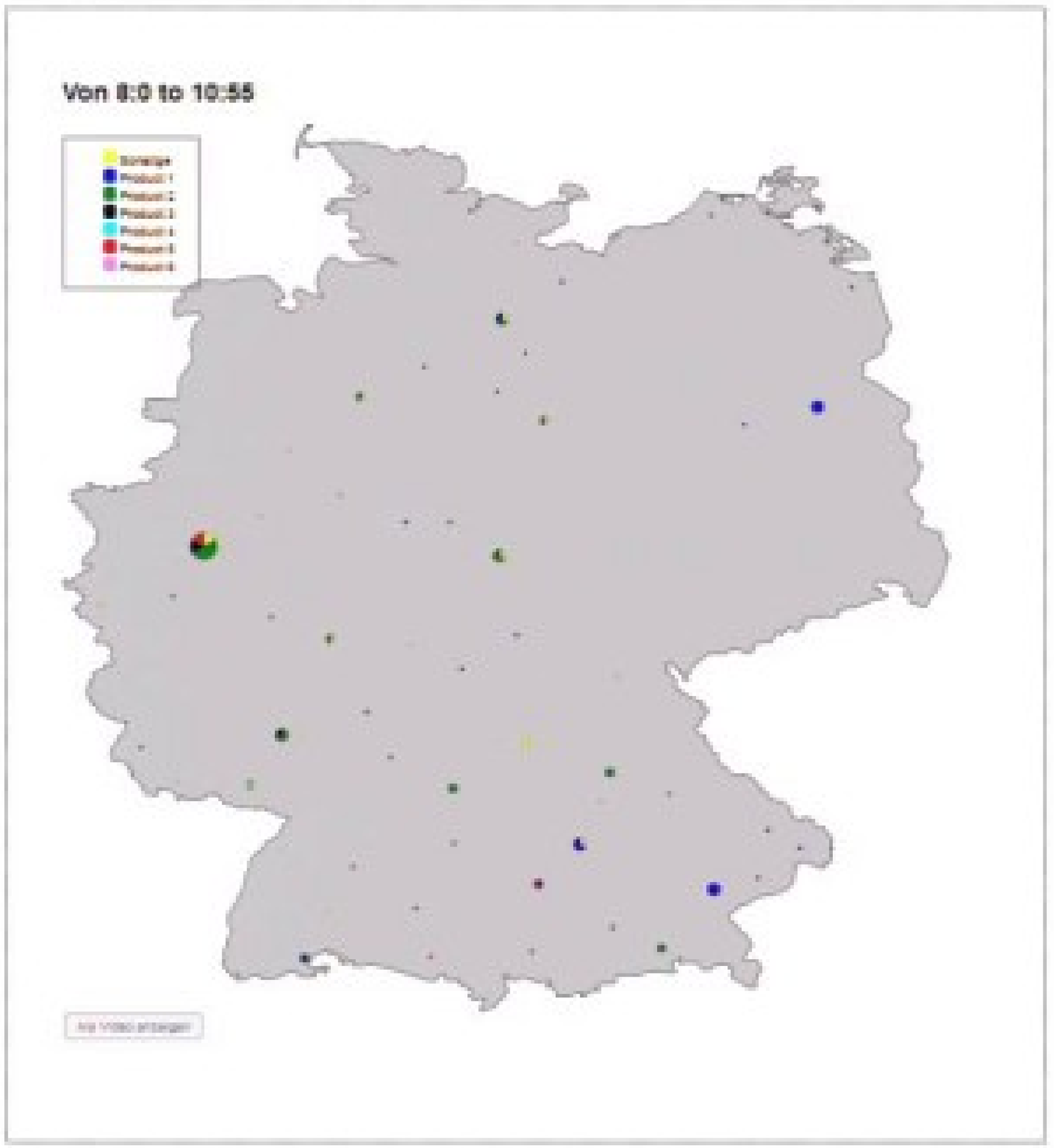


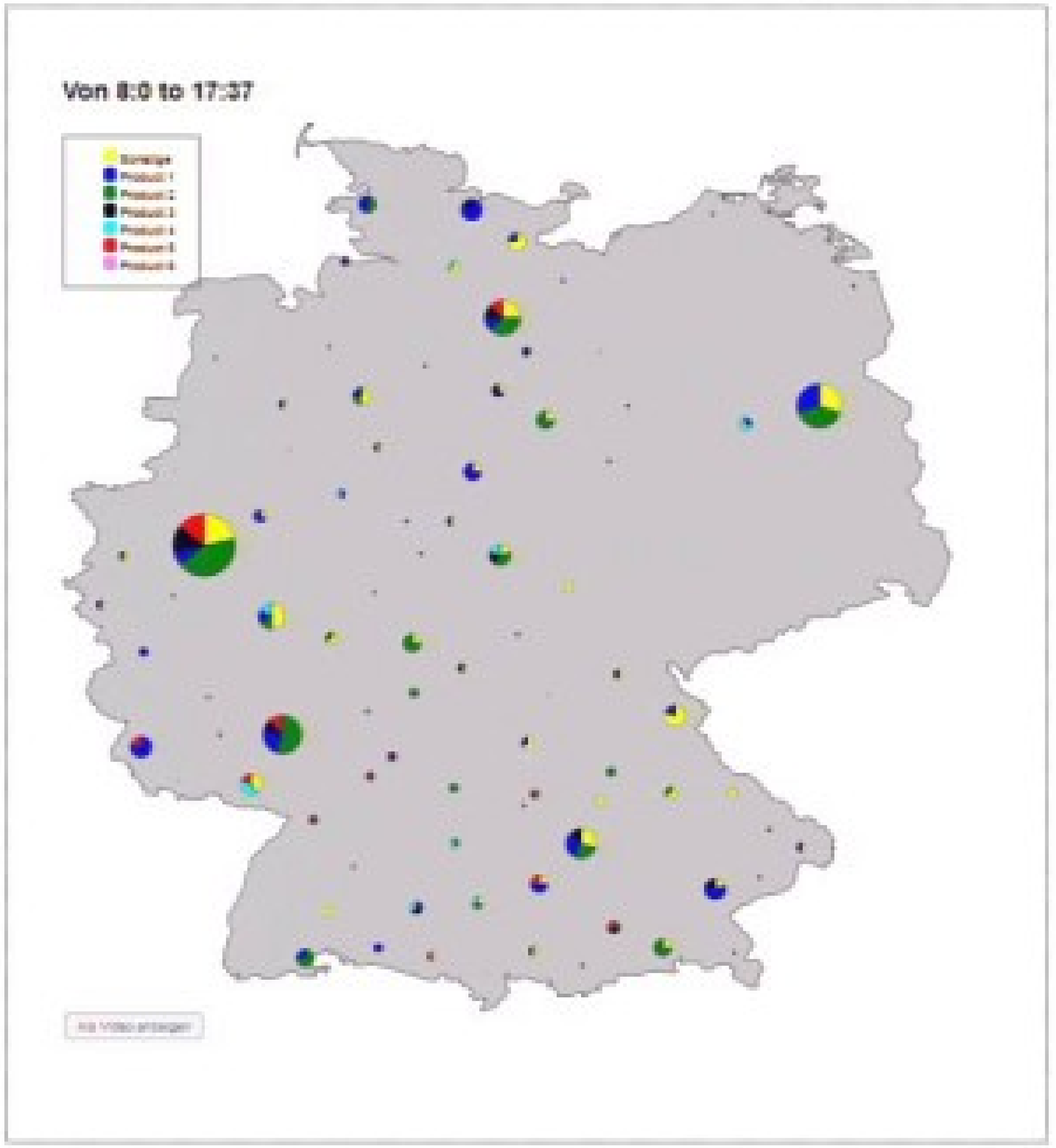

Von %from to %to
0:03
1:03

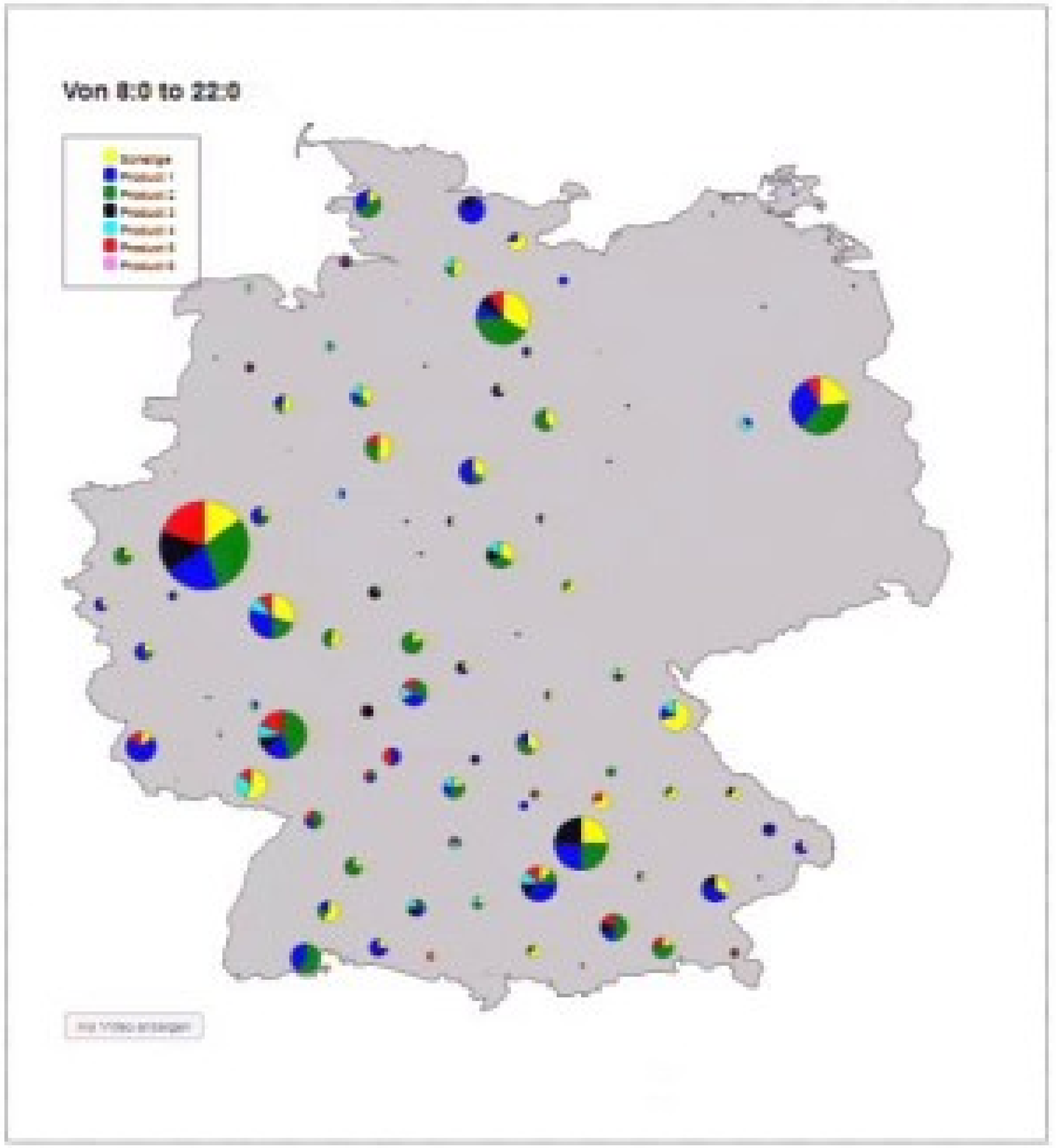
TNG TECHNOLOGY CONSULTING

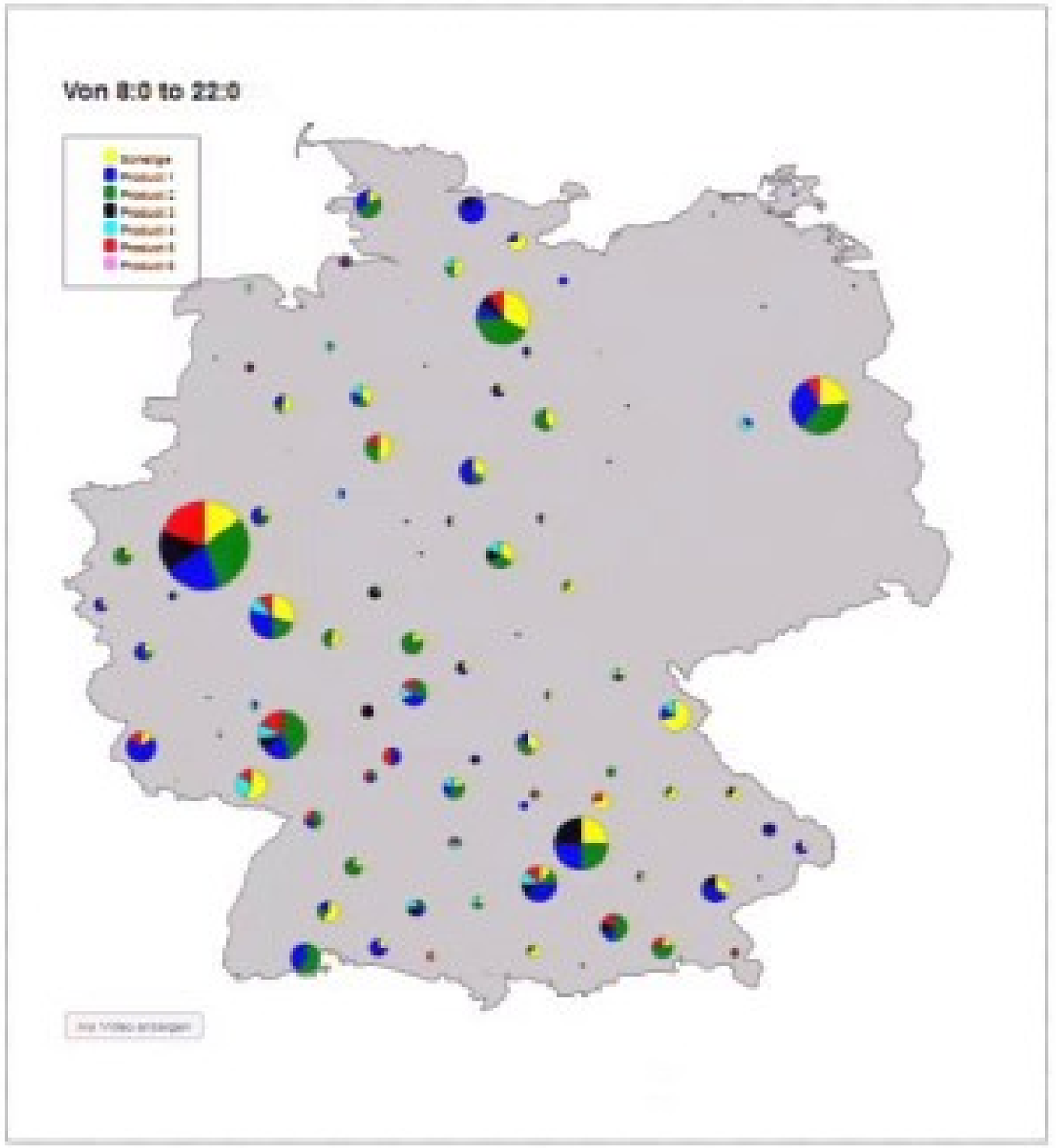

Von %from to %to

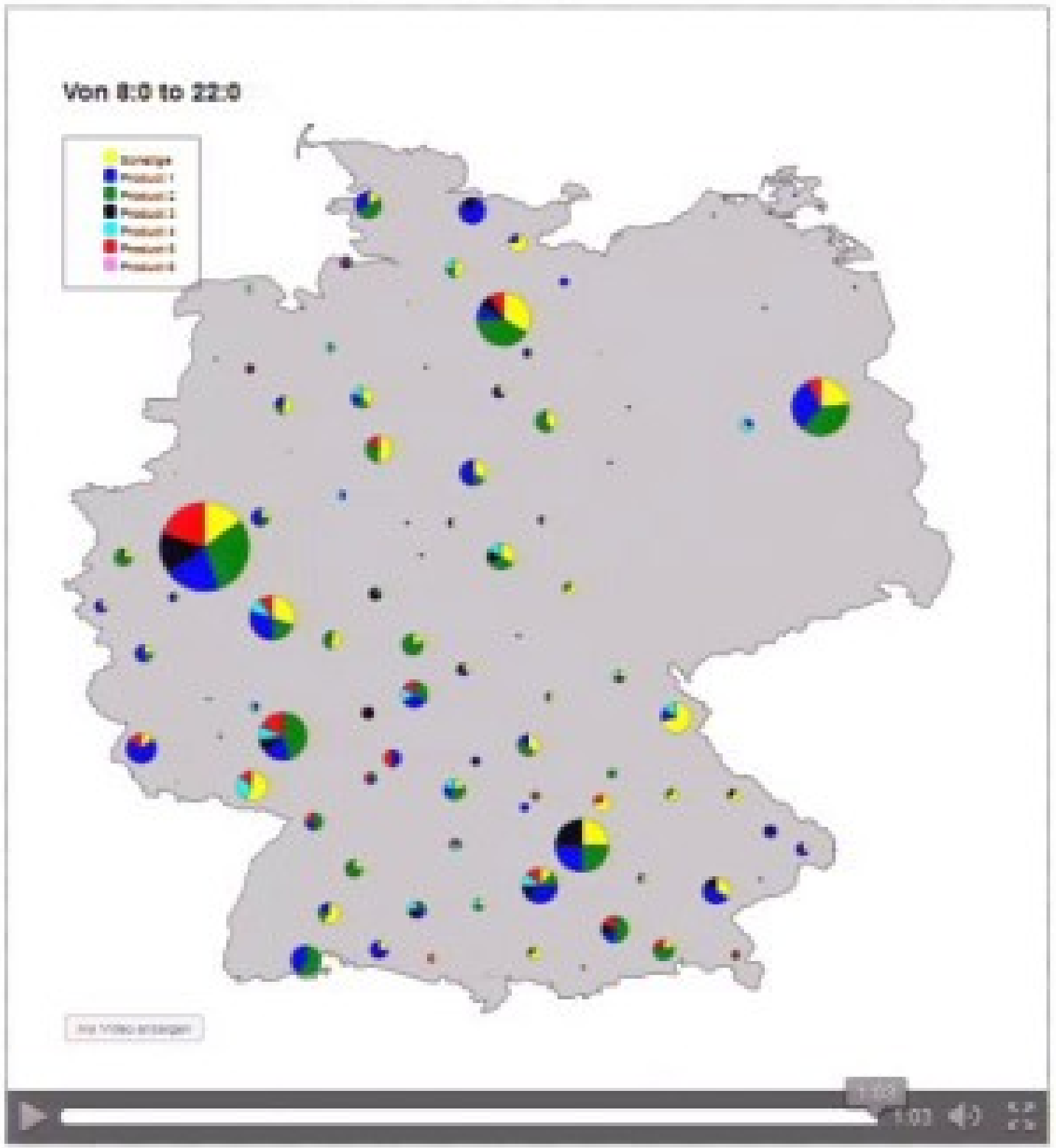

Von 8:0 to 8:17


Von 8:0 to 10:55


Von 8:0 to 17:37


Von 8:0 to 22:0

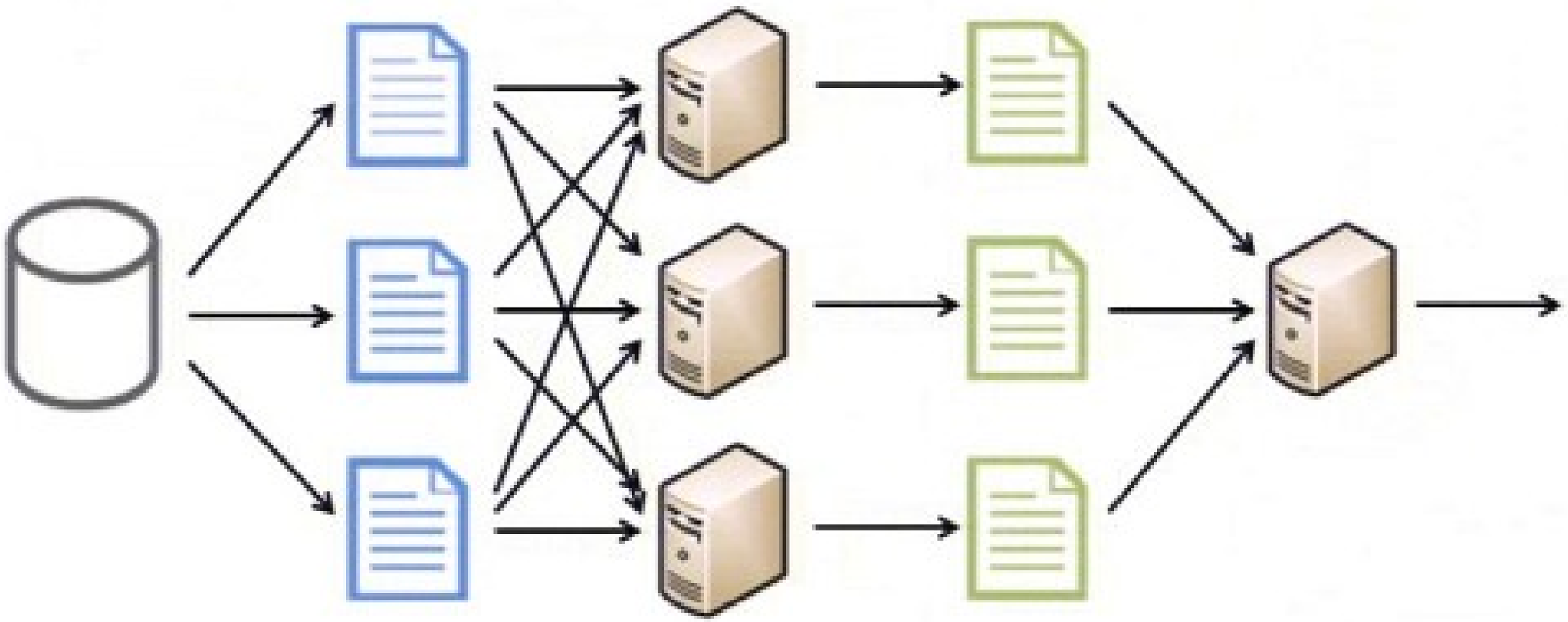

Von 8:0 to 22:0


Von 8:0 to 22:0
1:03

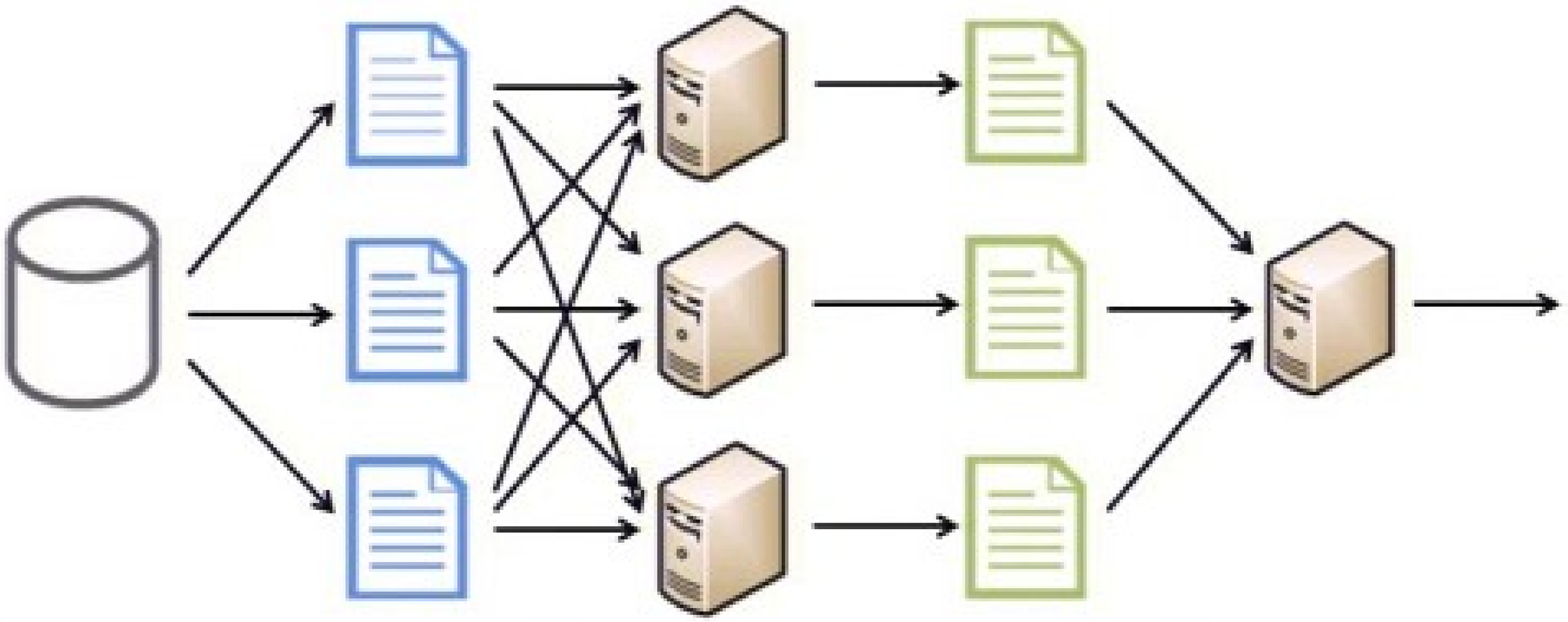
Challenges during implementation

- Map reduce is best suited for batch processing
- Integration of further data sources difficult
- Twitter and client portal did not provide useful data

Map / Reduce

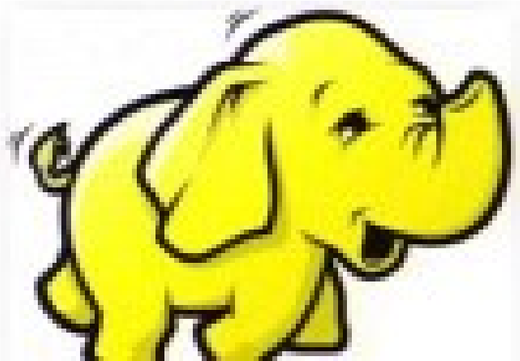


Map / Reduce



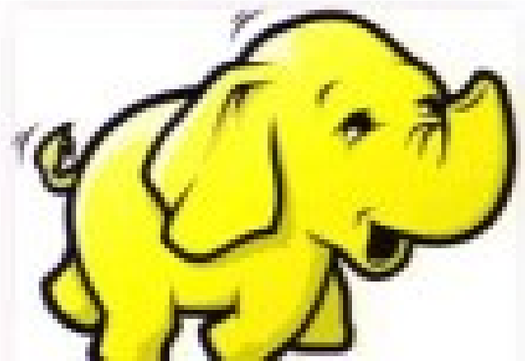
Hadoop: Distribute data storage and analysis

- Most successful implementation of map/reduce
- Provides infrastructure to scale across nodes
- Both data storage and computation tasks are distributed
- Hadoop API must be used to describe analysis jobs
- Commercial support: Cloudera, Hortonworks, ...



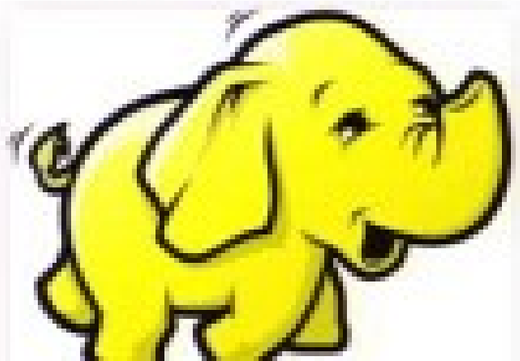
Hadoop: Distribute data storage and analysis

- Most successful implementation of map/reduce
- Provides infrastructure to scale across nodes
- Both data storage and computation tasks are distributed
- Hadoop API must be used to describe analysis jobs
- Commercial support: Cloudera, Hortonworks, ...



Hadoop: Distribute data storage and analysis

- Most successful implementation of map/reduce
- Provides infrastructure to scale across nodes
- Both data storage and computation tasks are distributed
- Hadoop API must be used to describe analysis jobs
- Commercial support: Cloudera, Hortonworks, ...



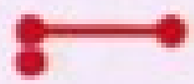
Hadoop: Components and sub-projects

- HDFS: Distributed file system for data storage
- MapReduce: Implentation of map/reduce algorithm
- Hive: Data mining Library which converts SQL queries to MapReduce
- Pig: High level data analysis language
- HBase: Transactional key/value store
- ... and ca. 10 more

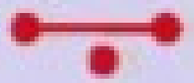
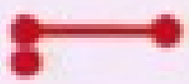
Realtime data processing frameworks

- Esper
- Storm Project

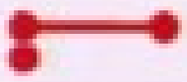
Temporal reasoning in stream processing

	Point-Point	Point-Interval	Interval-Interval
A before B			
A meets B			
A overlaps B			
A finishes B			
A includes B			
A starts B			
A coincides B			

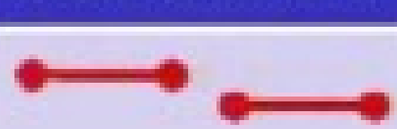
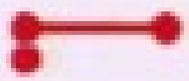
Temporal reasoning in stream processing

	Point-Point	Point-Interval	Interval-Interval
A before B			
A meets B			
A overlaps B			
A finishes B			
A includes B			
A starts B			
A coincides B			

Temporal reasoning in stream processing

	Point-Point	Point-Interval	Interval-Interval
A before B			
A meets B			
A overlaps B			
A finishes B			
A includes B			
A starts B			
A coincides B			

Temporal reasoning in stream processing

	Point-Point	Point-Interval	Interval-Interval
A before B			
A meets B			
A overlaps B			
A finishes B			
A includes B			
A starts B			
A coincides B			

(No)SQL for Big Data

- Key/value store
- Document database
- Column database
- various more exotic DB flavors

Open source

- Cassandra (hybrid key/value and column DB, Hadoop integration)
- MonetDB (relational column DB)
- Redis (key/value store)
- MongoDB (document DB)
- Many more...

The big vendors

- IBM bought Netezza (2010)
- EMC bought Greenplum (2010)
- HP bought Vertica (2011)
- SAP HANA
- Amazon: EMR, S3, DynamoDB
- Oracle has no own big data software

node.js: Event-driven data processing

- Server-side JavaScript
- Non-blocking IO
- Event-driven

Usecase 2: Broadband service management

- Analyze session accounting data from DSL access network
- Customer support for individual lines
- Identify regional outages
- Idea: Use multiple stages of map/reduce to correlate data
- Data protection requirements: no public clouds!

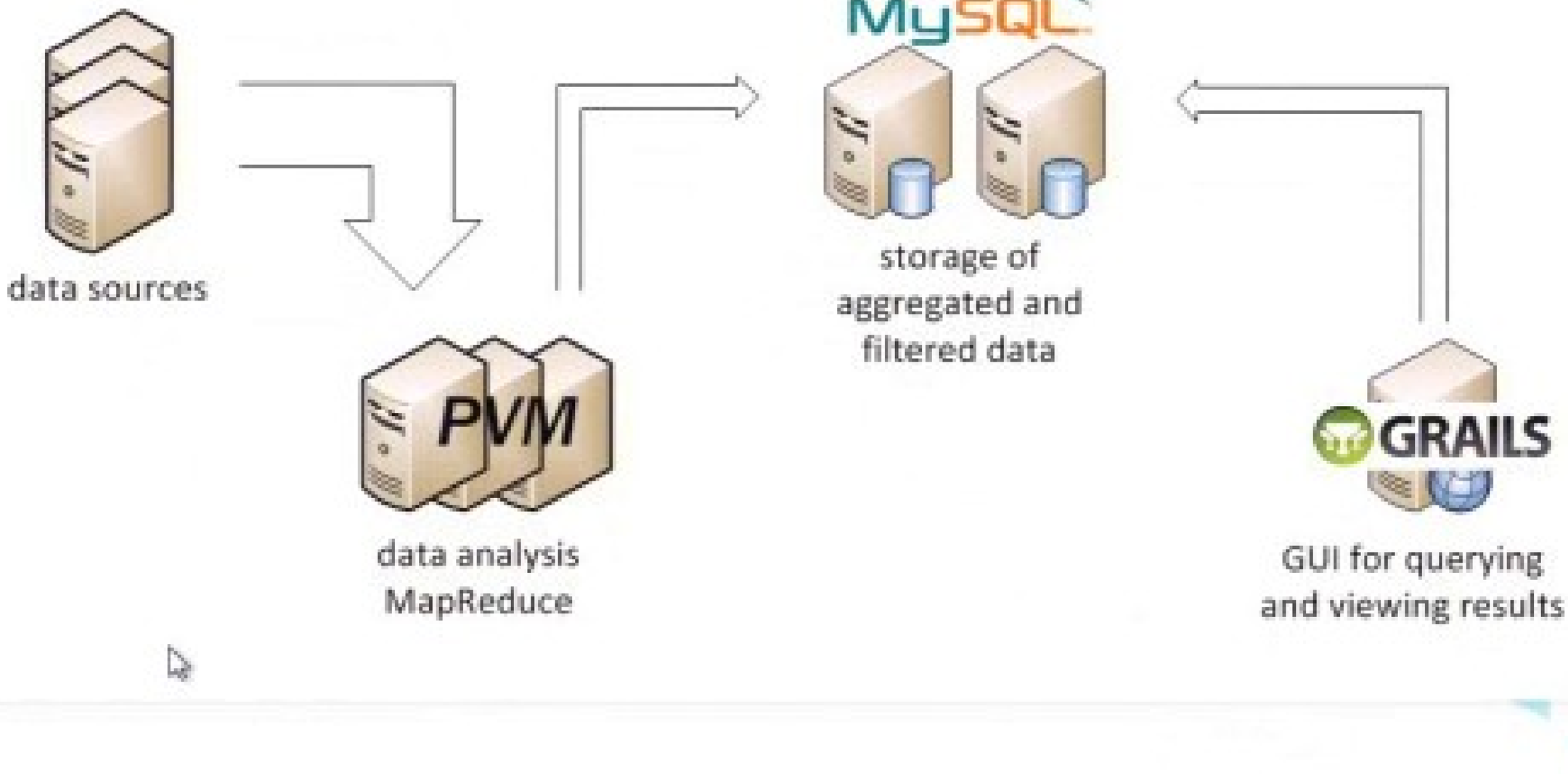
Performance figures

- Up to 10.000 records / second
- Approx. 100 million session in DB
- All data is updated / replaced within 7 days
- Availability of aggregated results after 1 minute
- Instantaneous results in GUI

Performance figures

- Up to 10.000 records / second
- Approx. 100 million session in DB
- All data is updated / replaced within 7 days
- Availability of aggregated results after 1 minute
- Instantaneous results in GUI

Application flow



Results

- Number and length of support calls are reduced
- Network outages can be located faster
- Faulty DSL routers are swapped proactively

Relational vs. NoSQL data store

- Classical relational DB allows easier enterprise integration
- NoSQL databases are more flexible
- Relational column-oriented databases try to combine the best of both worlds

Relational vs. NoSQL data store

- Classical relational DB allows easier enterprise integration
- NoSQL databases are more flexible
- Relational column-oriented databases try to combine the best of both worlds

Moving data is difficult

- Moving / copying data costs time
- Bandwidth is often a bottleneck
- Aggregate early to reduce the amount of data
- Move computing resources instead of data

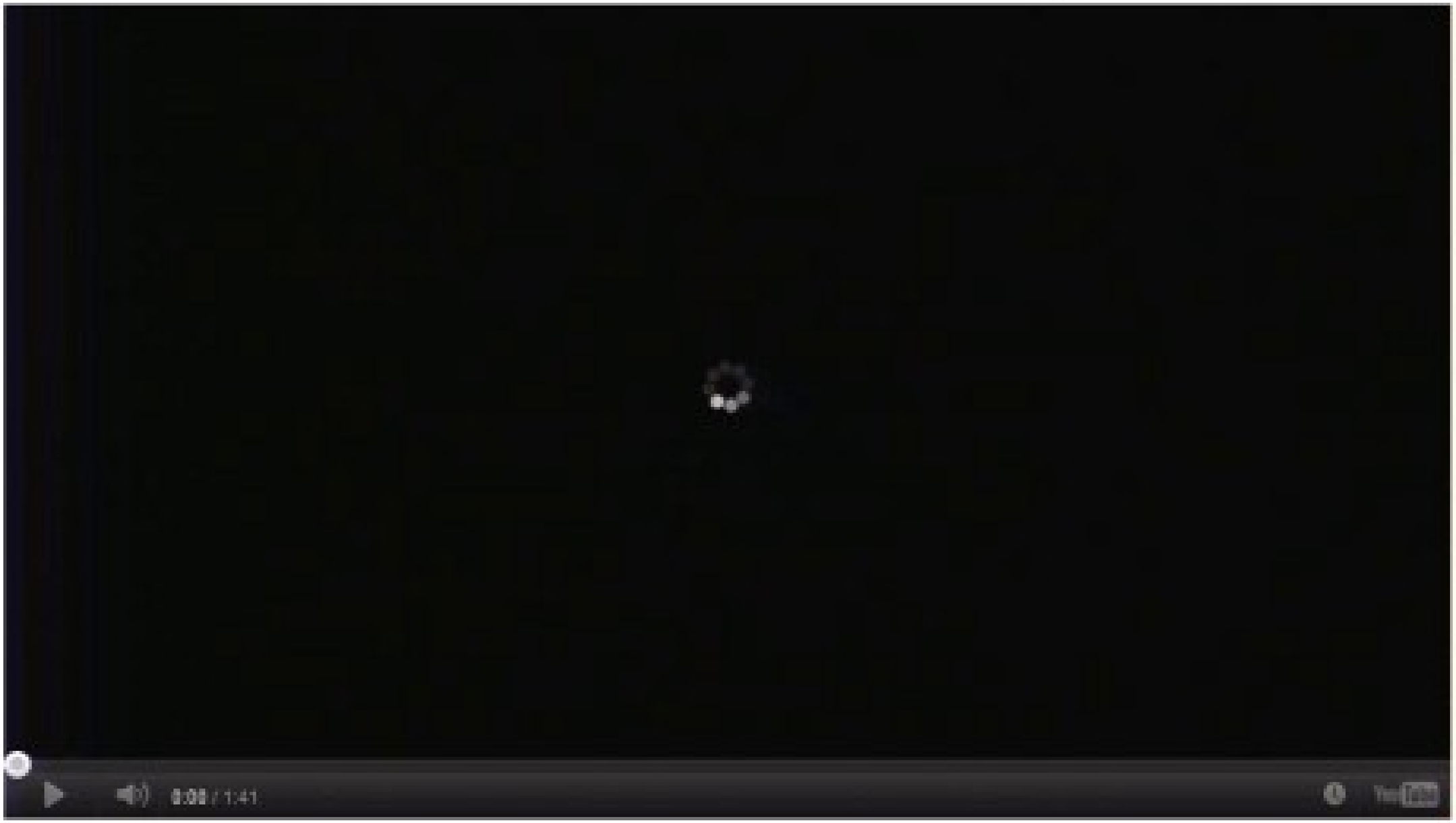
Mechanical sympathy

- Abstraction costs performance
- Embracing the architecture of your data sources and infrastructure will make you faster
- EMR, S3: Amazon takes care of that ⇒ limited API
- Compute clouds: Be careful how you use them!

Data source: Movement profile

source:Malte Spitz/YouTube

Data source: Movement profile



source: Malte Spitz / YouTube

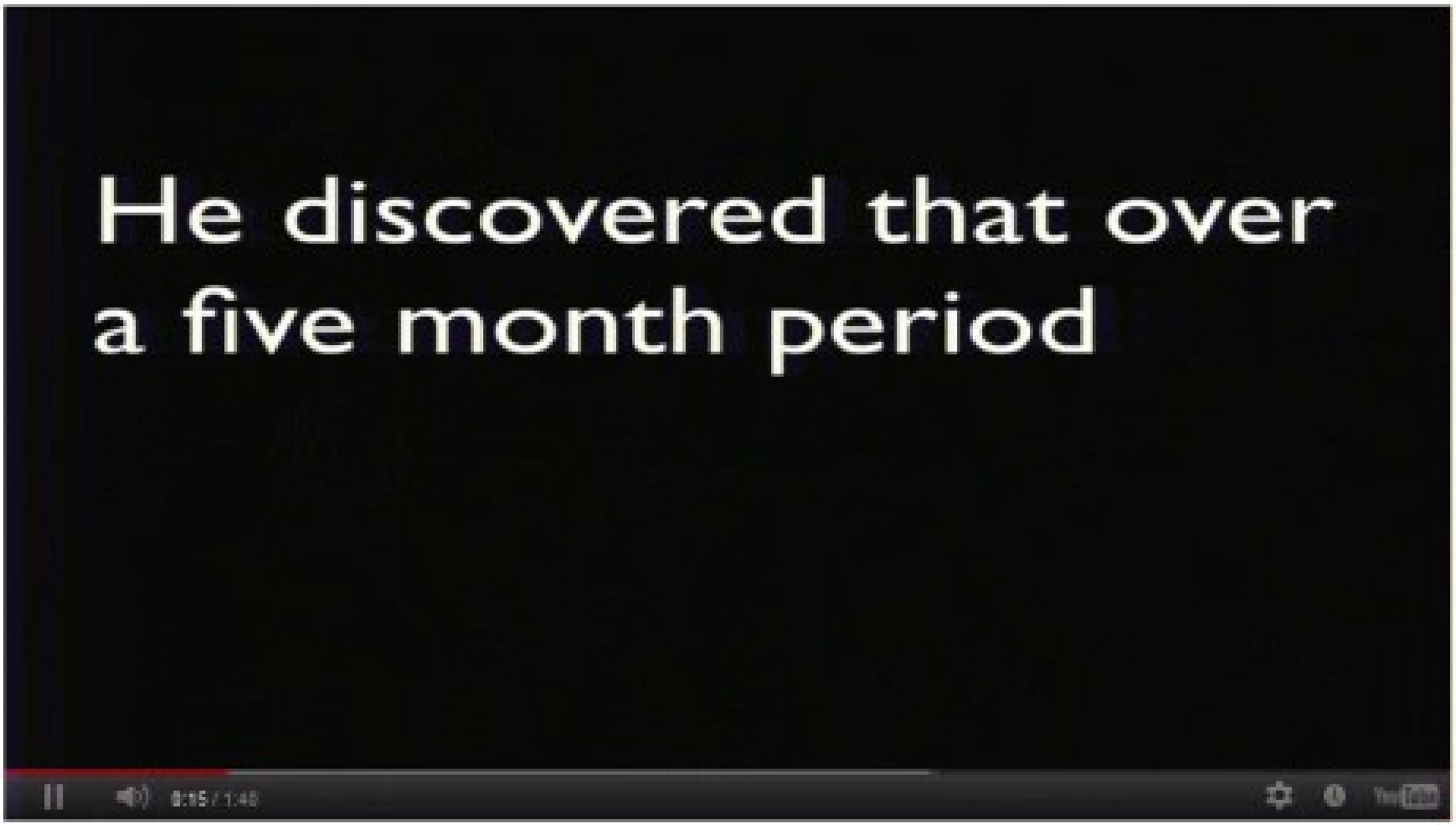
Data source: Movement profile

source: Malte Spitz / YouTube

Data source: Movement profile

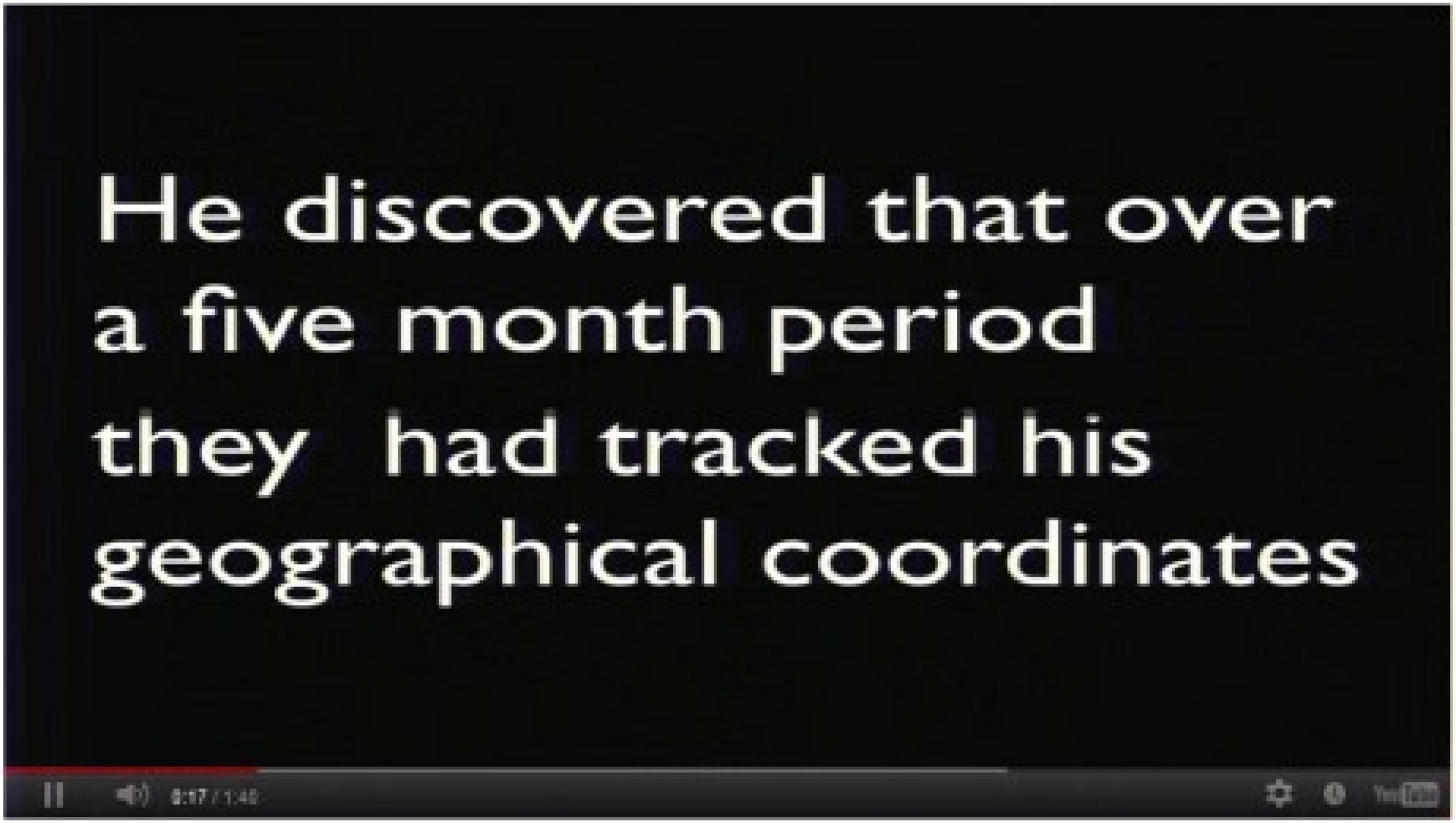
source: Malte Spitz / YouTube

Data source: Movement profile



source: Malte Spitz / YouTube

Data source: Movement profile



source: Malte Spitz / YouTube

Data source: Movement profile

source: Malte Spitz / YouTube

Data source: Movement profile

source: Malte Spitz / YouTube

Data source: Movement profile

source:Malte Spitz/YouTube

Use cases: Movement profile

- Traffic jam detection
- Moving money

Use cases: Movement profile

- Traffic jam detection
- Moving money

Use your existing data

- to improve your service
- to create new revenue / business

by applying the available Big Data technologies.

Have fun!

Documentation & Credits

- Thanks to Malte Spitz / YouTube
- Presentation done with reveal.js