positium

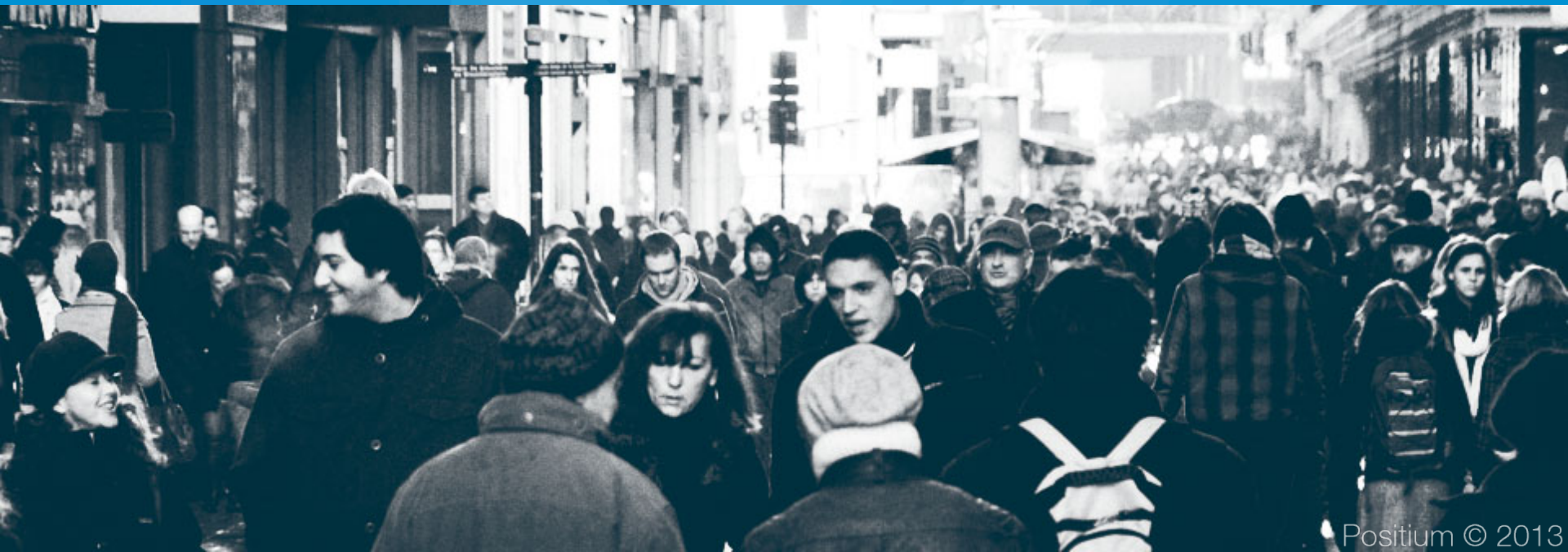
# Mobile Positioning Data Processing

Case Study 1

Aare Puussaar

Positium LBS

**15.10.2013**



# Agenda

- Little Intro
- Why am I here?
- Big Data
- Choosing the right solution
- Appling the technology with business logic
- What else can be done?

positium

Providing meaningful information service about location, movement flows and population statistics of humans

# What is mobile positioning?

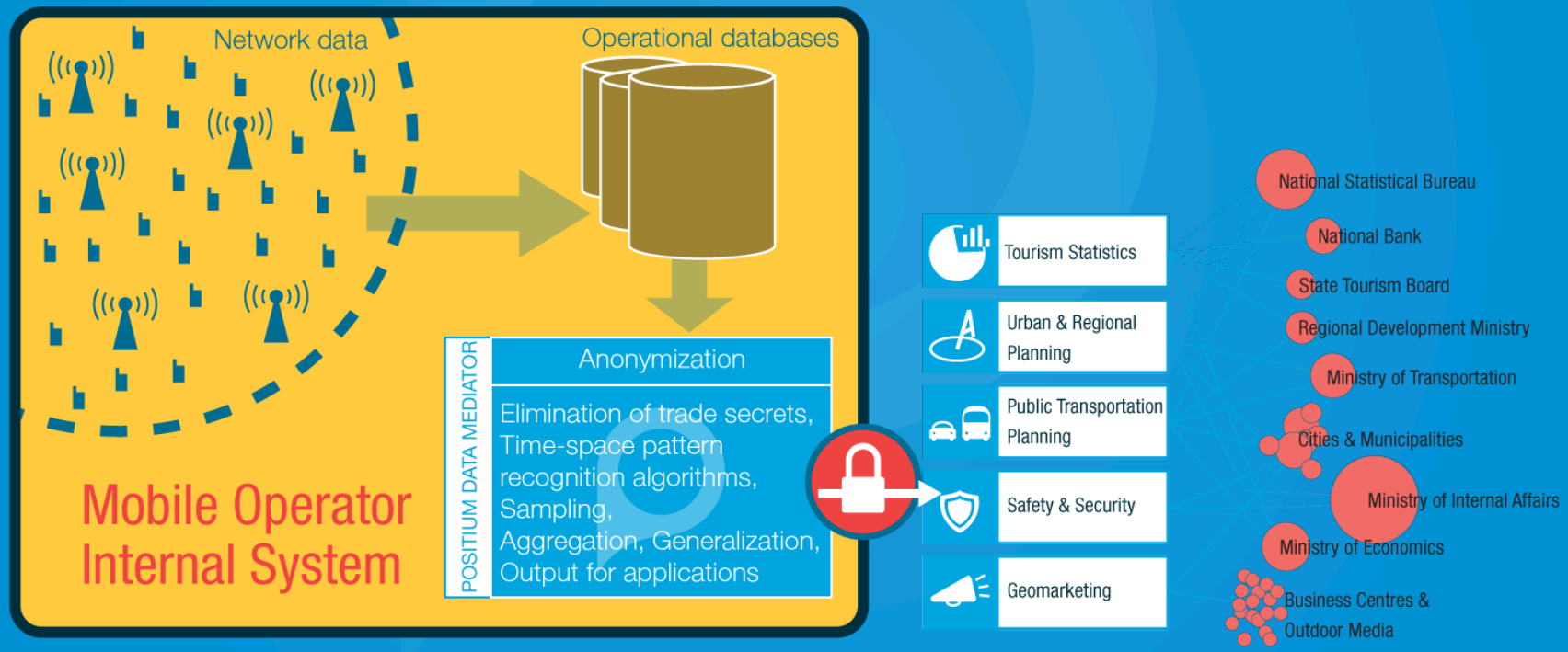
- Obtaining location of a mobile phone in time and space using cellular network and device position determining technologies
- Active (mostly identified)
  - Device-based
  - Network based (MPS)
- Passive (mostly pseudonymous)

# Anonymous (Pseudonymous)

## Passive positioning

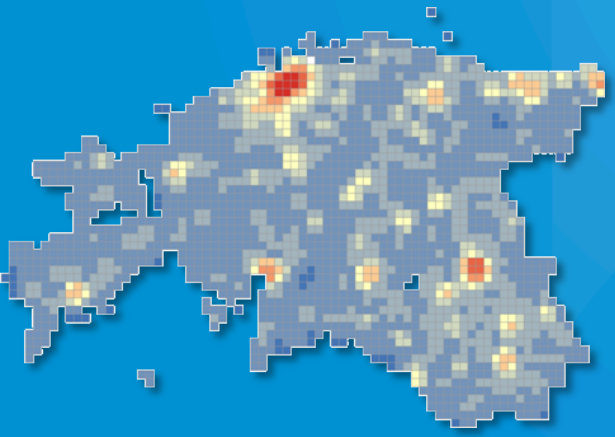
- Data originates from different registries, log-files of operator
- Call activity records (CDR), location area changes, data usage...

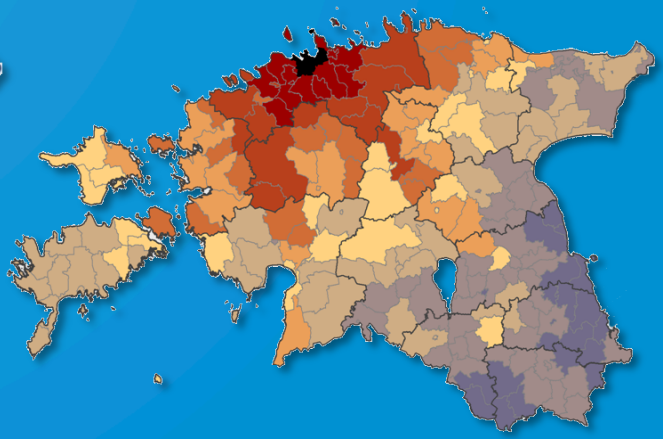
# Passive mobile positioning



# 24 hrs of data…

…holds more geographical information about movement than any register





# Motivation

- MTAT.08.027 Basics of Cloud Computing
- MTAT.03.280 Mobile and Cloud Computing Seminar
- We can do better…
- Professional interest

# What is Big Data?

The hot IT buzzword of 2012

Exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures.

**Big Data Now: 2012 Edition** by O'Reilly Media, Inc

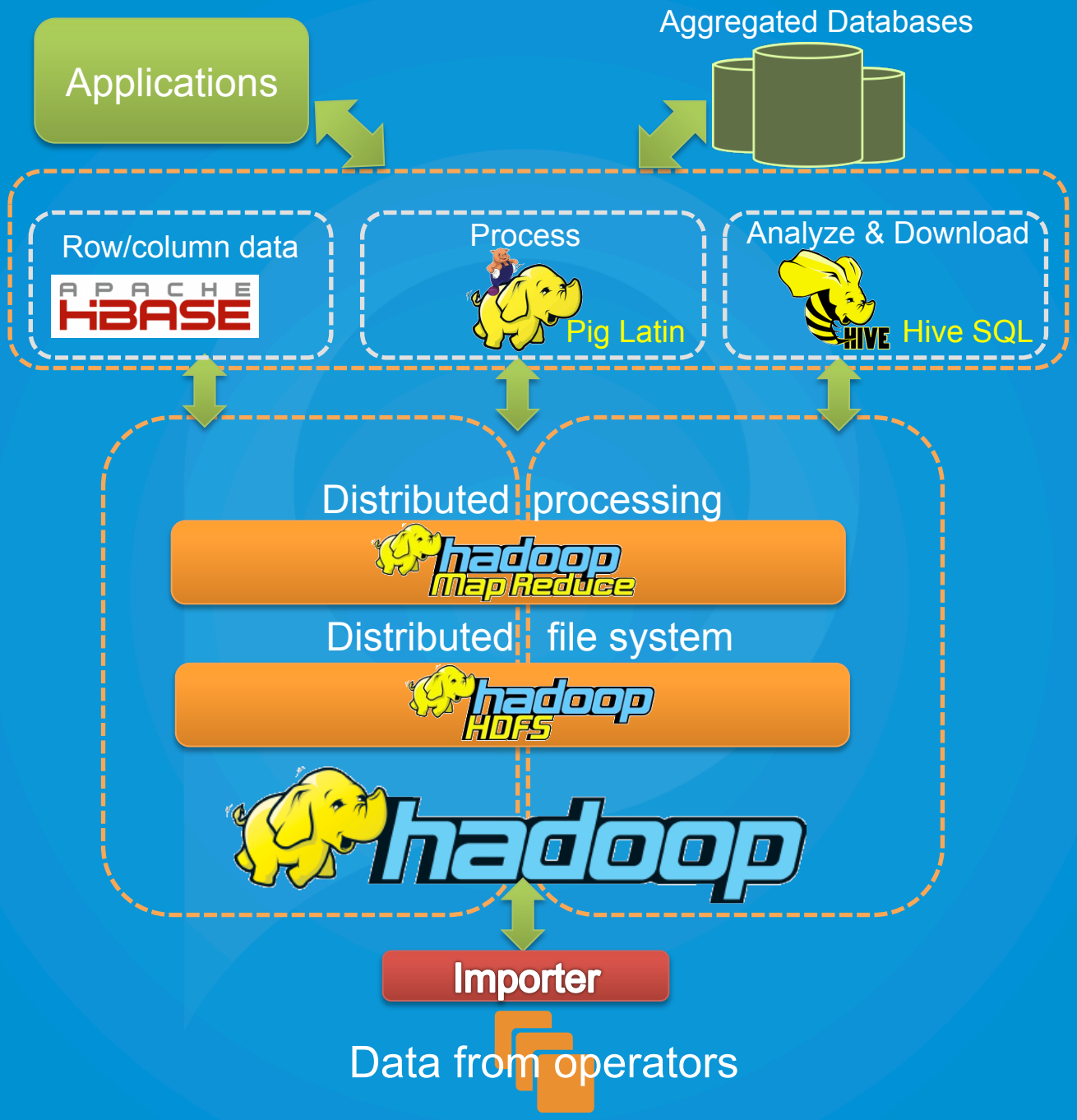
# Challenges with Big Data

- A lot of work for extracting value
- Power vs. Cost
- Speed
- Scalability


# The Technology – Use Case

# What is the right Solution?

- Cloud or own hardware
  - Privacy
  - Scalability
  - Cost
- More hardware
  - RAM
  - CPU
  - HDD
- Best stack for data



# Additions to stack

- Sqoop – “SQL to Hadoop”
  - Connectivity tool for moving data from relational databases and data warehouses into Hadoop.
- Oozie
  - A workflow engine and scheduler built specifically for large-scale job orchestration on a Hadoop cluster.
- In-Memory Accelerator For Hadoop®
  - In-memory HDFS and in-memory MapReduce

# What are the areas of research

- Urban geography, development, activity spaces, planning
- Transportation, traffic
- Seasonality of human spatial behaviour
- Tourism geography
- Spatial marketing (geomarketing)

# What are the areas of research Part 2

- Safety & security
- Environmental planning, climate change
- Ecologycal footprint
- Human mobility and psychology
- Genetics
- Epidemiology
- Privacy protection, legislation

# As a result of processing we can...

- Define home, work, free time anchors
- Everyday activity spaces
- Regular/non-regular trips
- Commuting
- Short-term migration (vacation trips, summer-houses)
- Long-term migration (change in home)

# Challanges

- Metadata must stay in memory while processing
- May need to update metadata while processing
- Processing generates huge intermediate data
- Large number of records and few aggregation groups
- Solution running on small cluster
- Multiple outputs from one input

# Make it work better…

- Better setup and smarter processing
  - Choose the best tool for the operation
  - Find out what is the best configuration
  - Choose right amount of map/reduce slots
- Better algorithms
  - Filter early and often
  - Reuse not replicate
  - Choose correct data types

# Hue – Hadoop User Experience

- Open source Hadoop UI released under the Apache License, version 2.0.
- Developed by Cloudera
- Has SDK

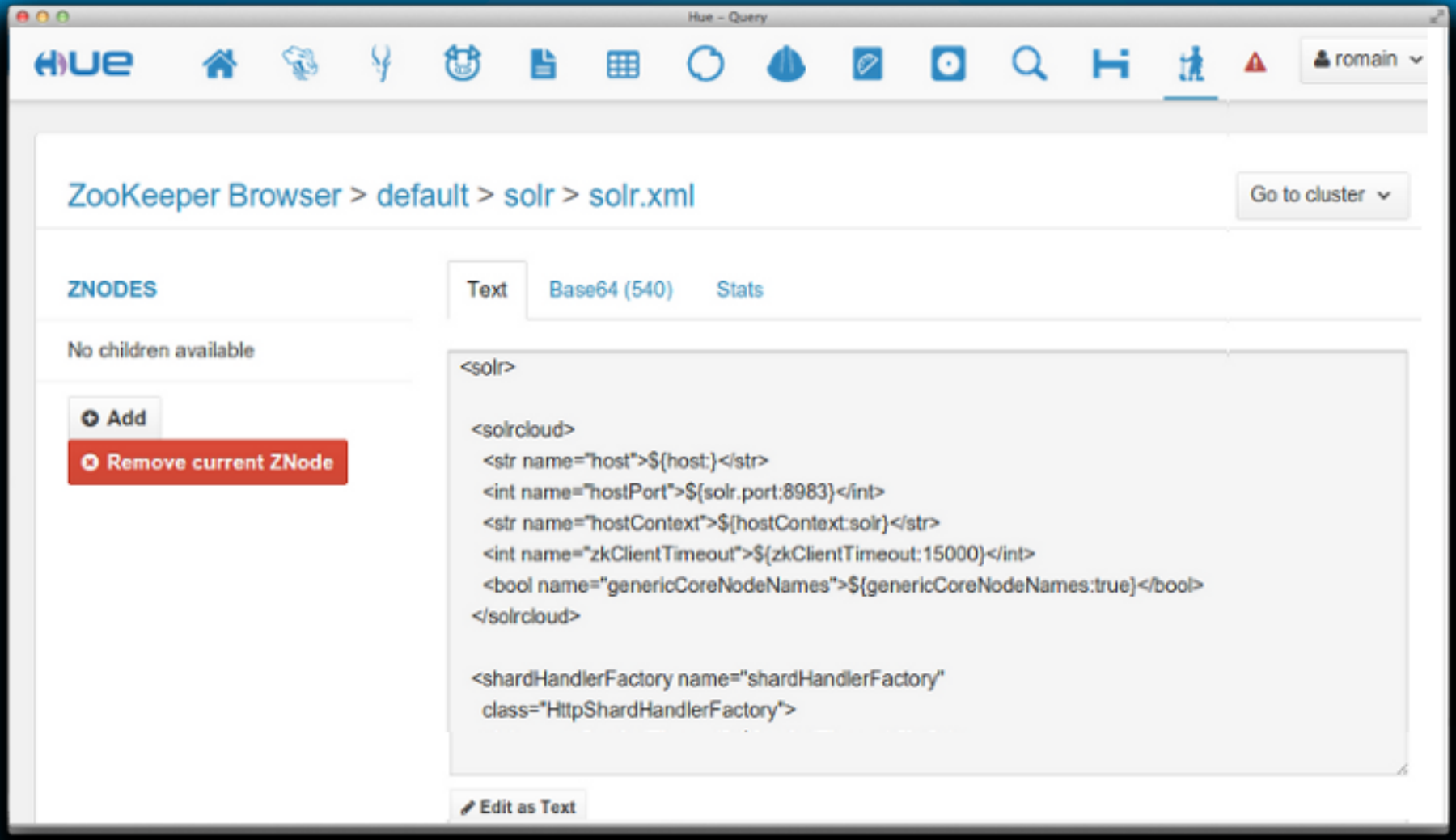
http://cloudera.github.io/hue/

# Hue features

- File Browser for HDFS
- Job Browser for MapReduce/YARN
- HBase Browser
- Query editors for Hive, Pig, Cloudera Impala and Sqoop2
- Oozie Application for creating and monitoring workflows
- Zookeeper Browser

http://cloudera.github.io/hue/

# Hue - Screenshots



# Literature

- **Tom White,** O'Reilly Media, Inc., **Hadoop: The Definitive Guide, Third Edition, May, 2012**
- Eric Sammer, O'Reilly Media, Inc., **Hadoop Operations, First Edition, Sept, 2012**
- Alan Gates, O'Reilly Media, Inc., **Programming Pig, First Edition, Oct, 2011**
- Edward Capriolo, Dean Wampler, Jason Rutherglen, O'Reilly Media, Inc., **Programming Hive, First Edition, Oct, 2012**

# Conclusions

- Data only grows
- Eventually everybody who is working with data runs into a wall
- Need smart system architects and tools

# Next Time

- Spark by Artjom Lind
  - Boost data analysis on cluster

# References


- O’Reilly Media, Inc., Big Data Now: 2012 Edition

positium

# Thank You for Attention!

Aare Puussaar
aare.puussaar@positium.ee

positium