Real-time streaming analytics of mobile phone data

Csaba Sidló
Institute for Computer Science and Control (SZTAKI),
Hungarian Academy of Sciences

Informatics Laboratory,
Business Intelligence and Data Warehousing Group

sidlo@sztaki.mta.hu

http://dms.sztaki.hu
http://bigdatabi.sztaki.hu

2013. június 4.

SZTAKI ILAB and Big Data

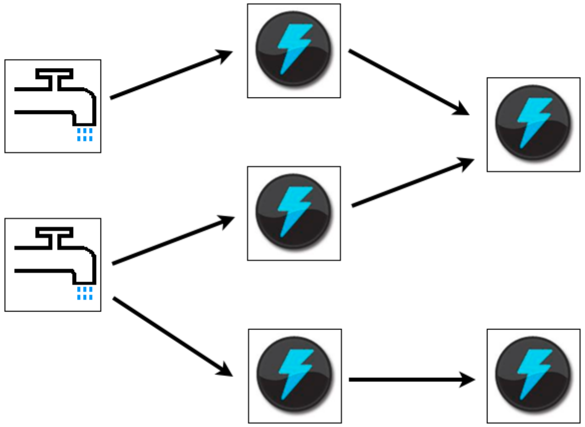
http://dms.sztaki.hu

- ILAB research groups:
  - András Benczúr, head, „Momentum” MTA grant on „Big Data” research
  - research and development – innovation, real-life applications
  - 30-40 members: researchers, developers, students
  - 60+ machines, 170+ cores, 600+ TB storage
- Big Data Business Intelligence Group
  - partner: Laboratory on Engineering & Management Intelligence , Dr. Zsolt János Viharos
- projects with „big data” problems
  - web- and log-analytics, web search, spam- and fraud-detection, recommender systems
  - smart city, mobility, „internet of things”

http://bigdatabi.sztaki.hu

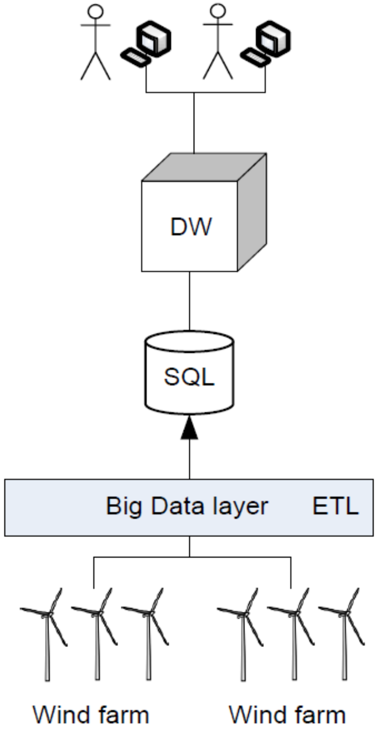
Interesting research topics

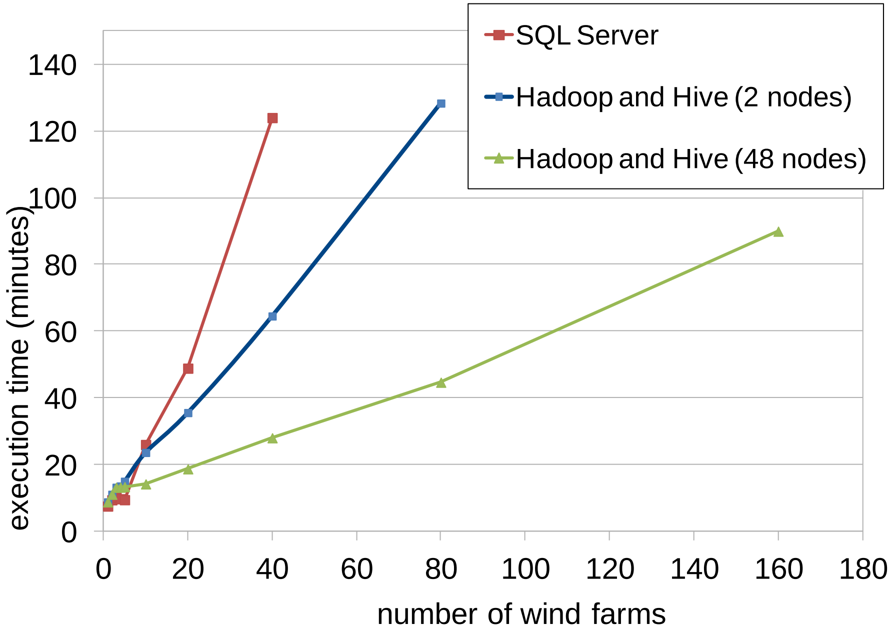
- IEEE BigData 2013?
- cloud, privacy, data integration,
  search and data mining eg. large scale graph processing, crowdsourcing, Internet of Things (Internet of Everything!), mobility,…
- scalable data management in a cloud:
  - storage systems: how to hide data locality,
    eg. multiple data canters and local computation in a cloud
- new computation models:
  - what is the next big thing after Hadoop / MapReduce?
  - simplicity and speed
    vs. supporting complex operations



Application: sensor data

Zs.J.Viharos, Cs.I.Sidló, A.A. Benczúr, J.Csempesz, K.Kis, I.Petrás, A.Garzó: **"Big Data" Initiative as an IT Solution for improved Operation and Maintenance of Wind Turbines**



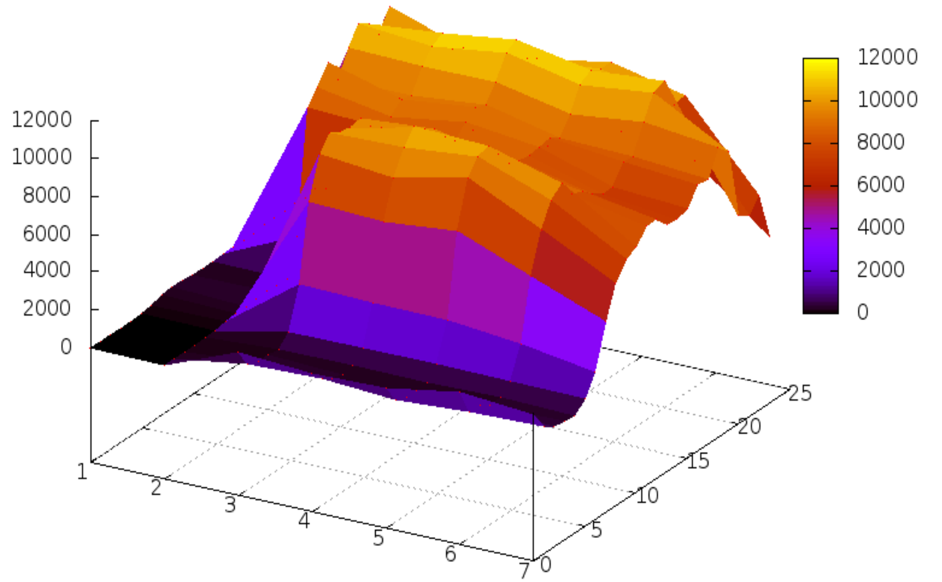


- experiments: wind farm data, substituting SQL DBs with Hadoop/Hive for handling most granular data
- efficient: sub-linear scalability, flexible, but **high latency**
- but maintenance requires **real-time, low-latency** alerts, statistics (high cost of maintenance)

Application: analysing mobile phone location data

- locating phones: at least cell tower granularity, when user is active
- opportunities:
  - anomaly detection, customer experience: improved service quality
  - smart city: traffic prediction, smart parking, bike hire schemes, optimize public transport
  - targeted ads, route optimization, city planning
  - detecting epidemic outbrakes, emergency situations
  - **low-latency** is required for lots of these applications
- difficulties:
  - hard to collect data beyond CDRs
  - custom data integration solutions
  - strict privacy constraints
  - no merged data sets of service providers

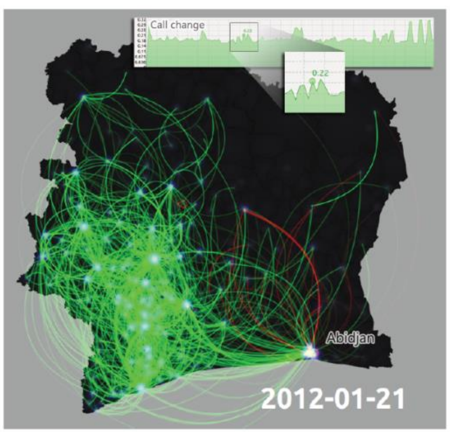
- “big data” competition open to the scientific community
    - exploring the tremendous potential of telephone data
    - producing rich, diverse ideas
- Orange anonymised data set: Ivory Coast, December 2011 → April 2012, ~ 5M users, 2.5 billion records
    - aggregate communication between cell towers
    - communication sub-graphs
    - mobility traces: privacy vs. fine resolution
        - coarse (prefectures) with more users,
        - **fine resolution dataset** with less users (sparse sample)



http://perso.uclouvain.be/vincent.blondel/netmob/2013/

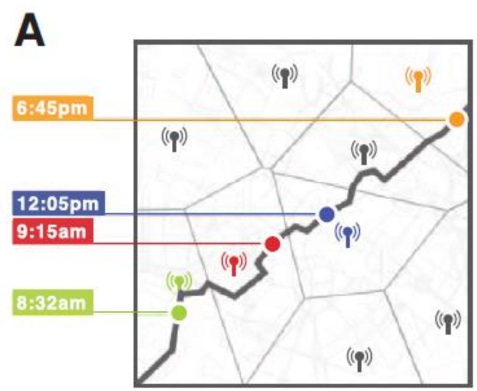
D4D main results

Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics approach

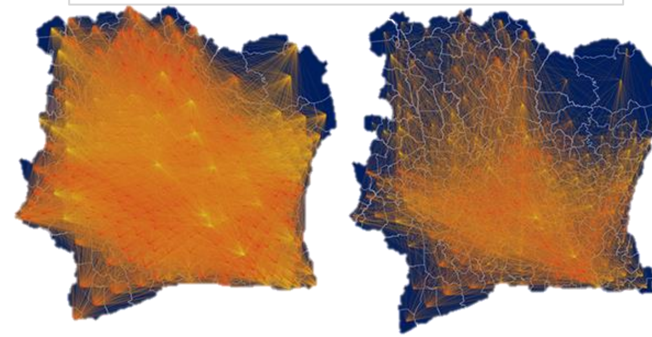


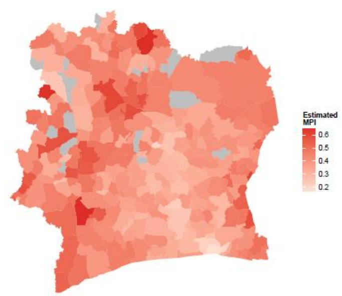
Unique in th crowd:

The privacy bounds of human mobility

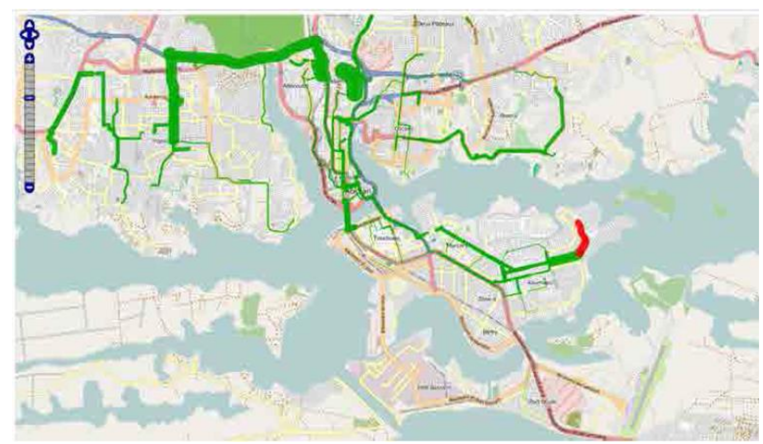


disease containment using calls matrix and mobility matrix

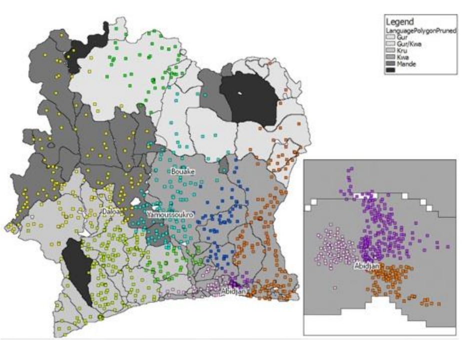




poverty map



AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data



Analyzing social divisions using cell phone data

Our goals

- predict user location → traffic
- with **real-time scalable distributed stream processing**
  100 000 events / sec
  (several million people)
- key research tasks:
  - scalability (horizontal, by increasing #servers)
  - real time response
  - fault tolerance (many commodity machines)
  - software layers to ease analytics development

MTA SZTAKI Hungarian Academy of Sciences Institute for Computer Science and Control

Which tools to choose?

Big Data Landscape

Vertical Apps

Predictive Policing, bloomreach, Myrrix

Log Data Apps

splunk, loggly, sumologic

Ad/Media Apps

rocketfuel, collective[i], bluefin, Recorded Future, Media Science, Turn, LuckySort, DataXu

Business Intelligence

Oracle | Hyperion, SAP BusinessObjects, RJMetrics, Microsoft Business Intelligence, IBM, Cognos, birst, Autonomy, MicroStrategy, QlikView, bime, Domo, Chart.io, GoodData

Analytics and Visualization

tableau, Palantir, Opera, metaLayer, Metamarkets, dataspora, Teradata Aster, centrifuge, sas, TIBCO, Karmasphere, panopticon, pentaho, Datameer, platfora, ClearStory, Cirro, alteryx, visual.ly, Ayata

Data As A Service

factual., kaggle, knoema beta, GNIP, DataSift, Windows Azure Marketplace, INRIX, LexisNexis, Space Curve, Loqate

Analytics Infrastructure

Hortonworks, Vertica, MapR, cloudera, Infobright, ParAccel, EMC², Greenplum, Netezza, kognitio, DataStax, Exasol, calpont

Operational Infrastructure

Couchbase, 10gen, Teradata, Hadapt, Terracotta, VoltDB, MarkLogic, Informatica

Infrastructure As A Service

amazon web services, Windows Azure, infochimps, Google BigQuery

Structured Databases

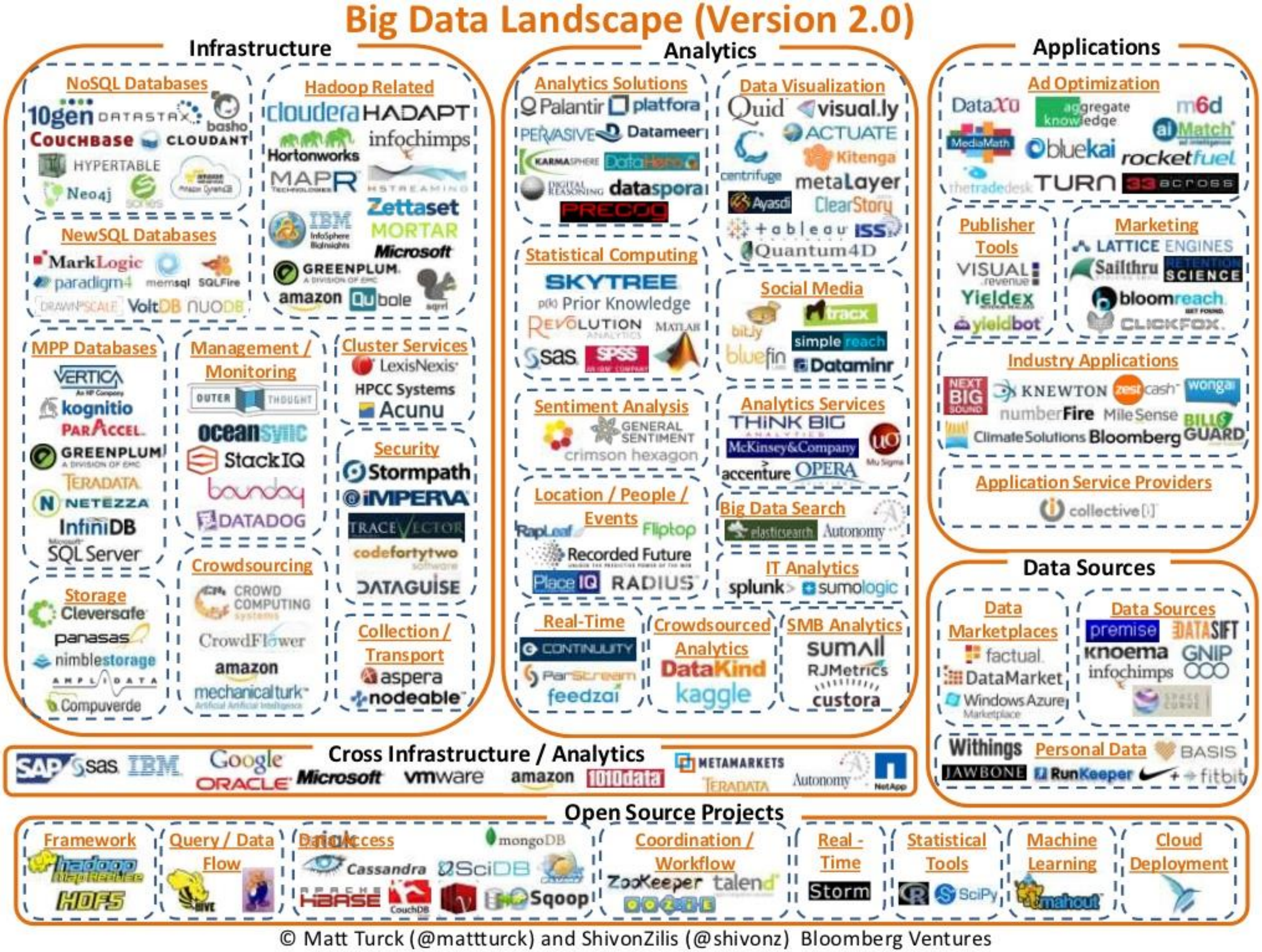
Oracle, MySQL, SQL Server, PostgreSQL, IBM DB2, Sybase, memsql

Technologies

hadoop, hadoop MapReduce, mahout, Apache HBase, Cassandra

 dave@vcdave.com blogs.forbes.com/davefeinleib

2012.06: http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/

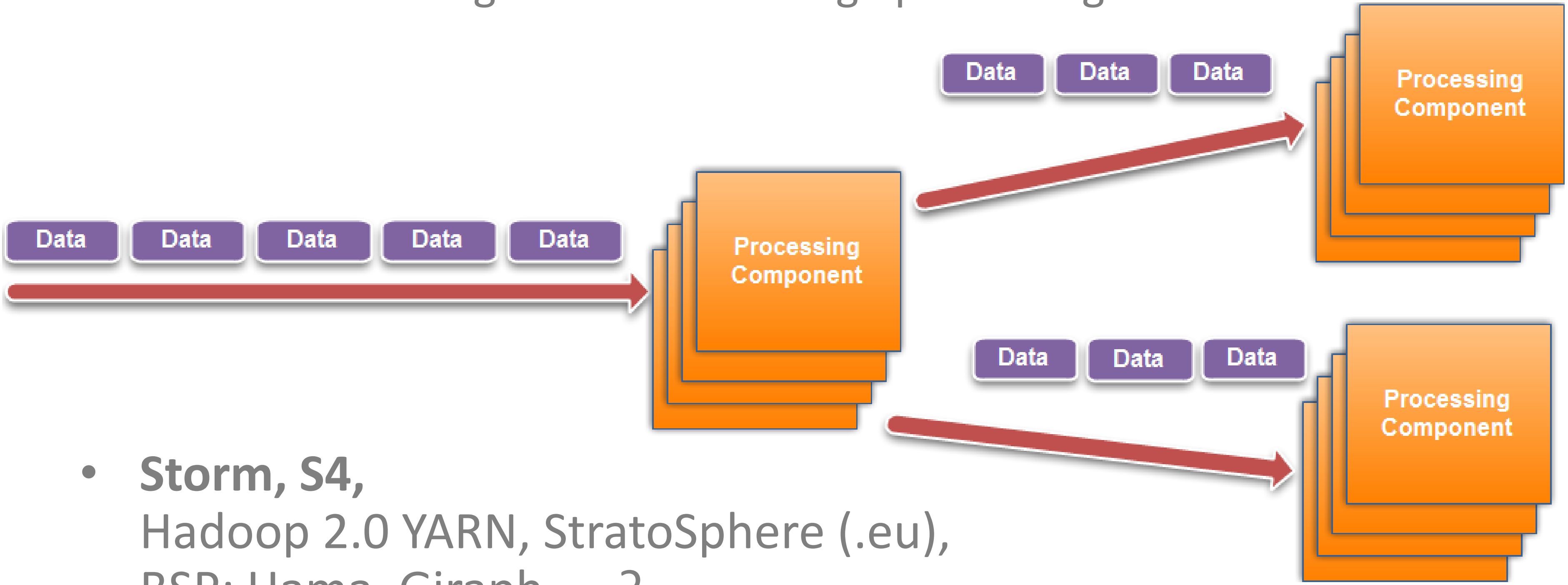


MTA SZTAKI Hungarian Academy of Sciences Institute for Computer Science and Control

2013.02: http://www.slideshare.net/mjft01/big-data-big-deal-a-big-data-101-presentation

Distributed stream processing tools

- distributed stream processing:
  - processing components run parallel
  - data passed by streams among components
  - acyclic execution graph can be defined by the user
  - nice to have: guaranteed message processing

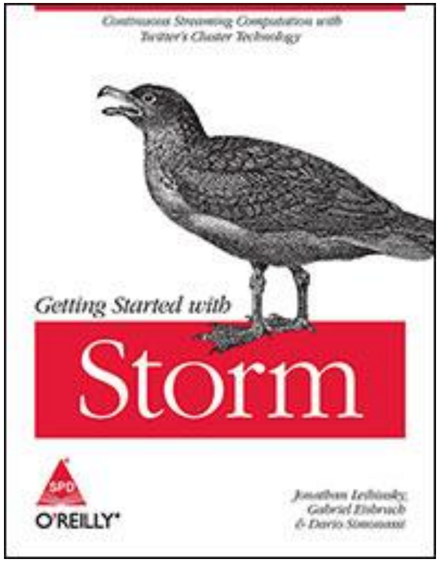


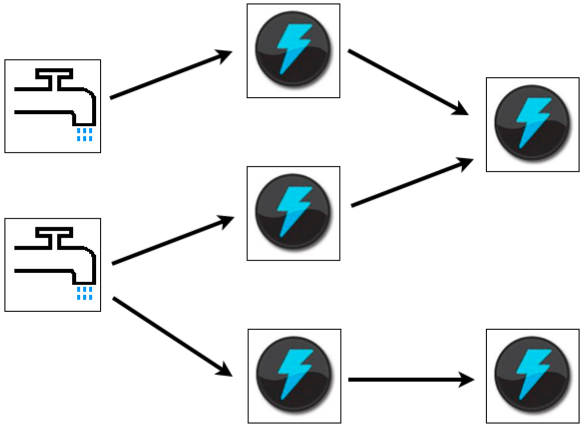
- **Storm, S4,**
  Hadoop 2.0 YARN, StratoSphere (.eu),
  BSP: Hama, Giraph, … ?

MTA SZTAKI Hungarian Academy of Sciences Institute for Computer Science and Control

Storm

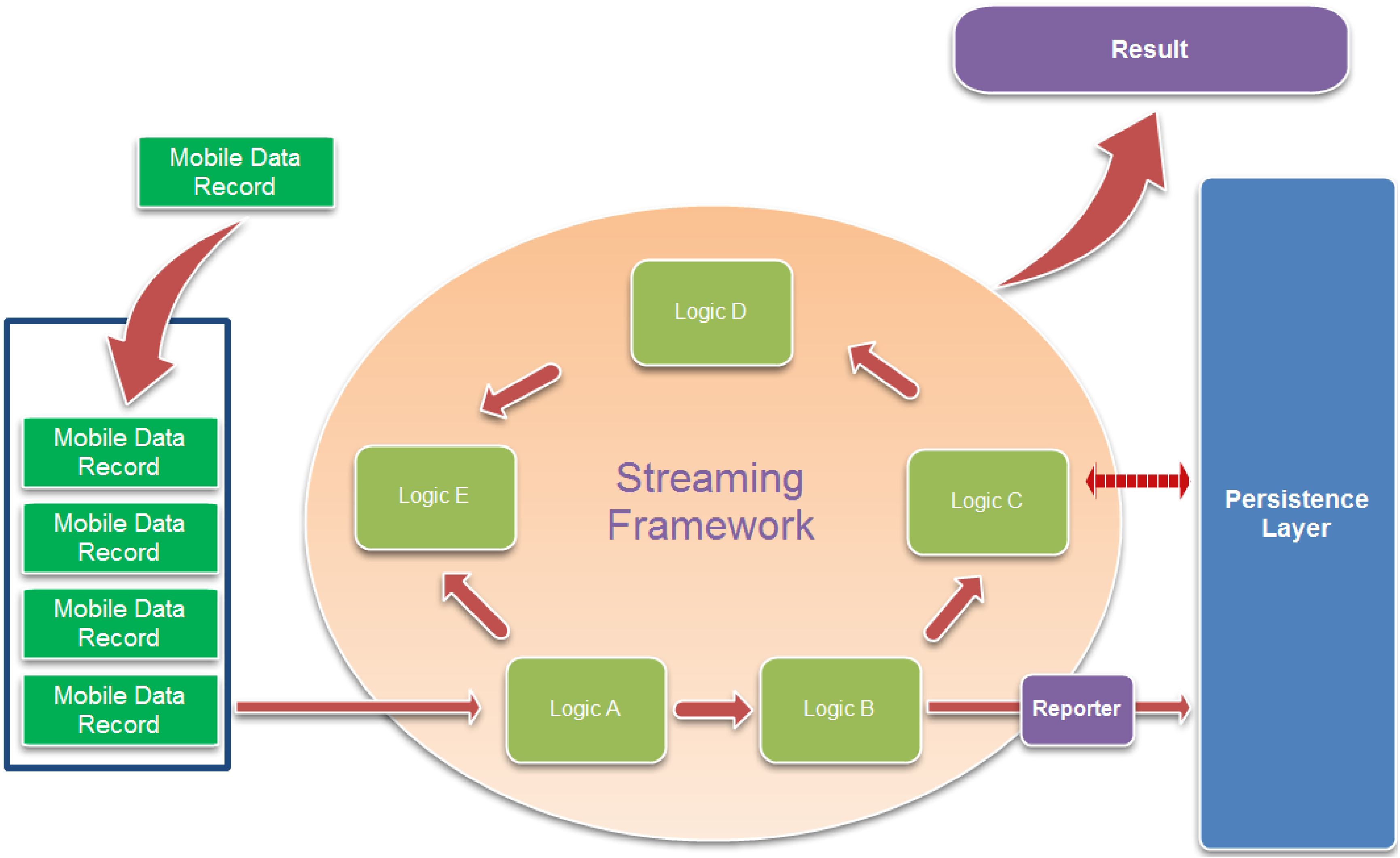
- guaranteed data processing
- horizontal scalability
- fault-tolerance
- no intermediate message brokers
- no single point of failure
- higher level abstraction than message passing
- “just works”,
  „Hadoop of real time streaming jobs”
- built by Backtype,
  recently bought by Twitter
- available as Open source
- Java + Closure,
  still under development
  (with an active community)



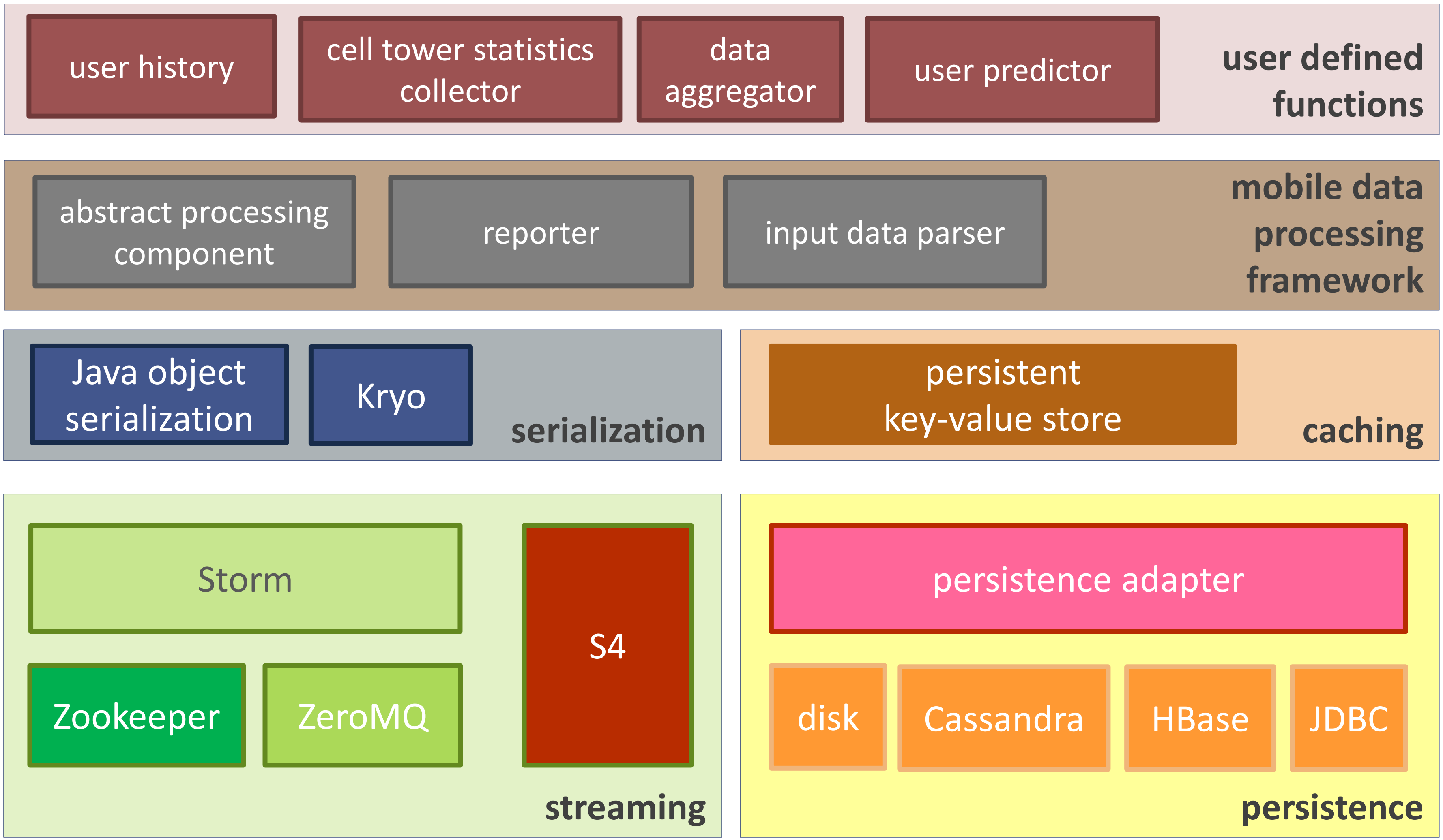


source: http://storm-project.net/

A framework for real-time prediction



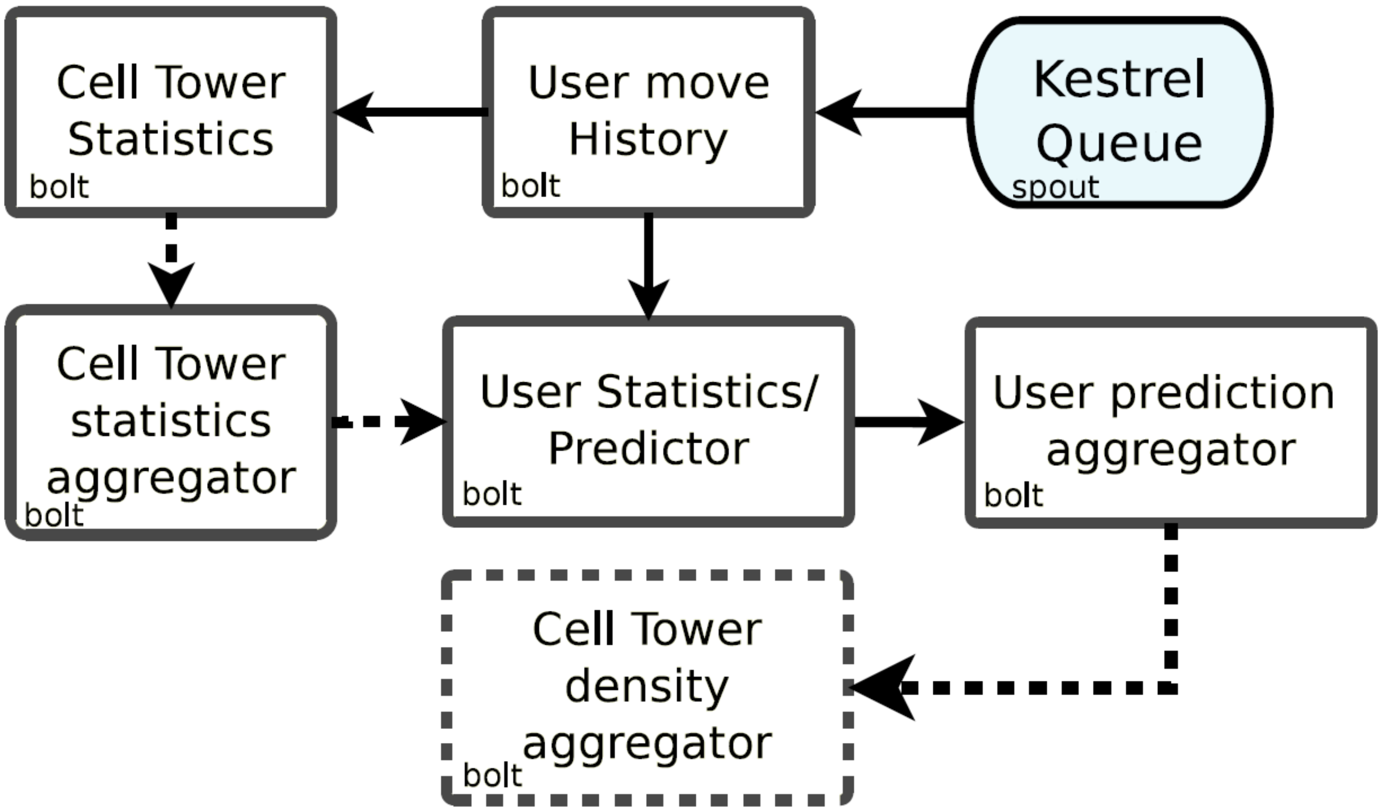
A framework for real-time prediction



MTA SZTAKI Hungarian Academy of Sciences Institute for Computer Science and Control

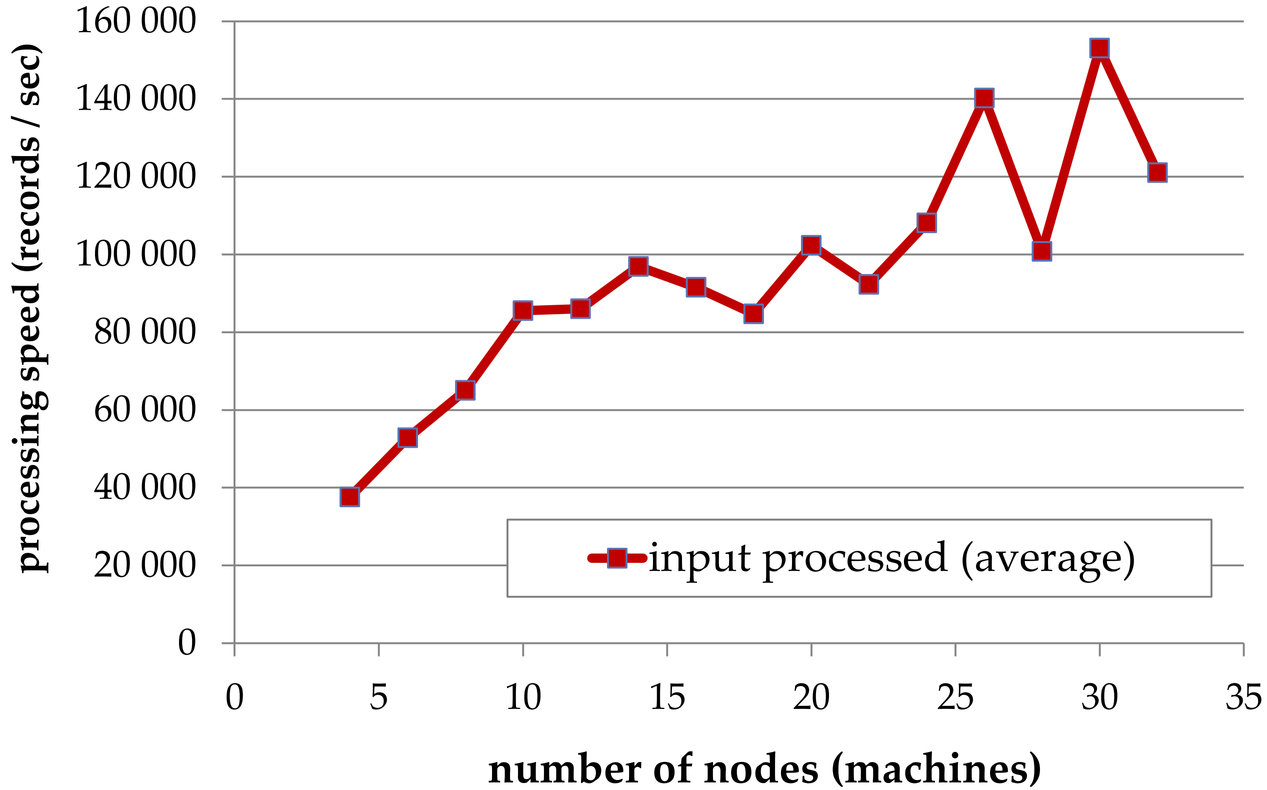
Processing components for prediction

- simple user and tower models for D4D:
  - discrete time intervals
  - tree of frequent paths, typical movement directions for cell towers

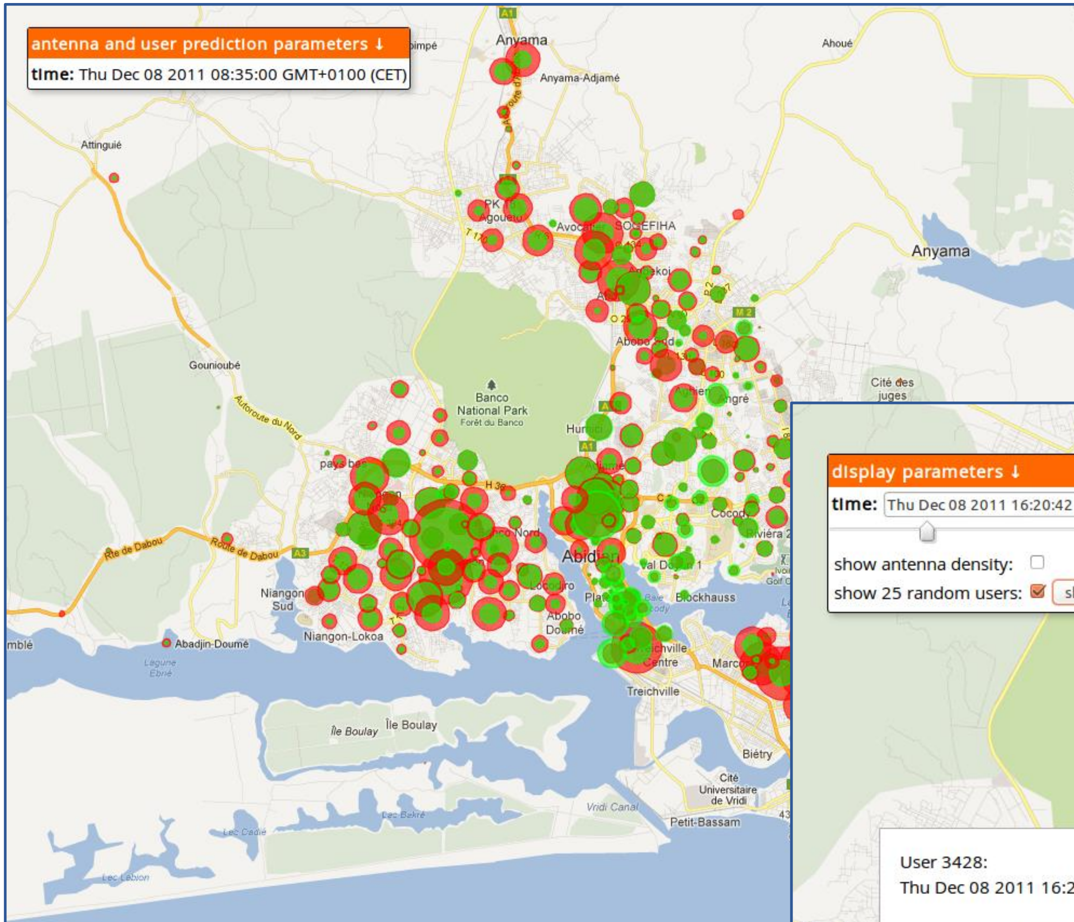


Experiments

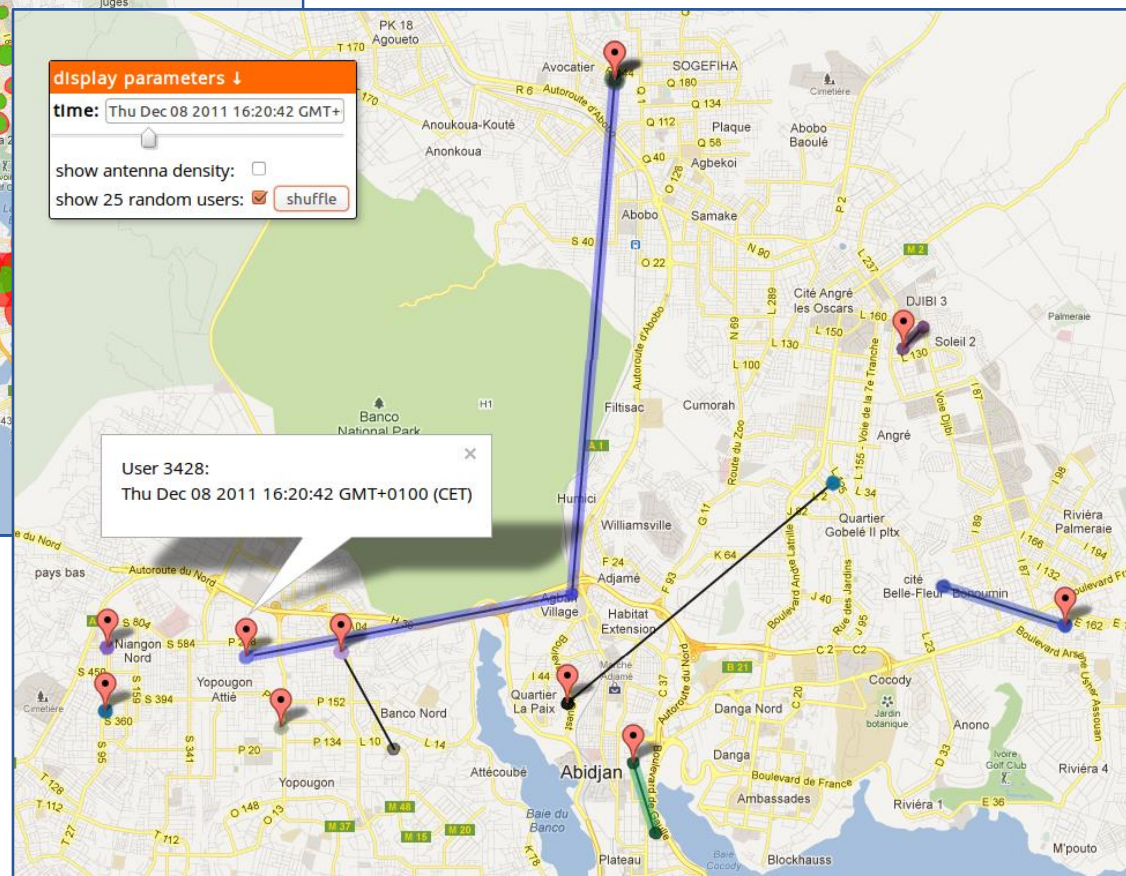
- Storm 0.9.0-wip4, old dual core Pentium-D 3GHz, 4GB machines
- with dynamic time warping, real location is predicted with 87.7% accuracy – most users just stay in place ☹
- latency: few seconds, <10
- recovery: depends on the persistence layer, but replaces a node within 10 min.



Demo visualization interface



aggregated cell density prediction



sample of individual user predictions

MTA SZTAKI Hungarian Academy of Sciences Institute for Computer Science and Control

Conclusions

- big data real-time analytics don’t have mature solutions yet
- but real-time location prediction is feasible on big data
- Storm is OK with some tricky parts which we still have to learn
- our framework lets machine learning guys do machine learning, and applicable to similar problems
- persistence layer can ensures fault tolerance

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

source: https:// flickr.com/photos/t_gregorius/5839399412