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

• ILAB research groups:
  o András Benczúr, head, „Momentum” MTA grant on „Big Data” research
  o research and development – innovation, real-life applications
  o 30-40 members: researchers, developers, students
  o 60+ machines, 170+ cores, 600+ TB storage

• Big Data Business Intelligence Group
  o partner: Laboratory on Engineering & Management Intelligence, Dr. Zsolt János Viharos

• projects with „big data” problems
  o web- and log-analytics, web search, spam- and fraud-detection, recommender systems
  o smart city, mobility, „internet of things”
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:
  o storage systems: how to hide data locality, eg. multiple data centers and local computation in a cloud
• new computation models:
  o what is the next big thing after Hadoop / MapReduce?
  o simplicity and speed vs. supporting complex operations
Application: sensor data

- 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)

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
Application: analysing mobile phone location data

• locating phones: at least cell tower granularity, when user is active

• opportunities:
  o anomaly detection, customer experience: improved service quality
  o smart city: traffic prediction, smart parking, bike hire schemes, optimize public transport
  o targeted ads, route optimization, city planning
  o detecting epidemic outbreaks, emergency situations
  o **low-latency** is required for lots of these applications

• difficulties:
  o hard to collect data beyond CDRs
  o custom data integration solutions
  o strict privacy constraints
  o no merged data sets of service providers
• “big data” competition open to the scientific community
  o exploring the tremendous potential of telephone data
  o producing rich, diverse ideas

• Orange anonymised data set: Ivory Coast, December 2011 → April 2012, ~ 5M users, 2.5 billion records
  o aggregate communication between cell towers
  o communication sub-graphs
  o 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 the crowd: The privacy bounds of human mobility

Analyzing social divisions using cell phone data

poverty map

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

disease containment using calls matrix and mobility matrix
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
Which tools to choose?
# Big Data Landscape

## Vertical Apps
- Predictive Policing
- Bloomreach
- MYRRX

## Log Data Apps
- Splunk
- Loggly
- Sumologic

## Ad/Media Apps
- RocketFuel
- Collective
- Recorded Future
- LuckySort
- MediaScience
- Turn
- Data2Go
- DataInsight

## Business Intelligence
- Oracle
- Hyperion
- SAP
- BusinessObjects
- Rimetrics
- Microsoft
- IBM
- Cognos
- Birt

## Analytics and Visualization
- Tableau
- Tableau
- Metamarkets
- Teradata
- Aster
- SAS
- Tibco
- Karmasphere
- Pentaho
- Pentaho
- Datameer
- Panopility
- Metabase
- ClearStory
- VisualLy
- Ayata

## Data As A Service
- Factual
- Gnop
- Datasift
- INRIX
- LexisNexis
- Kaggle
- Knome
- Locate

## Analytics Infrastructure
- Hortonworks
- Cloudera
- Vertica
- MapR
- ParAccel
- EMC
- Greenplum
- Netezza
- MarkLogic

## Operational Infrastructure
- Couchbase
- 10gen
- Teradata
- Hadoop
- Hadoop
- MarkLogic

## Infrastructure As A Service
- Amazon Web Services
- Windows Azure
- infochimps
- Google BigQuery

## Structured Databases
- Oracle
- MySQL
- SQL Server
- IBM
- DB2
- Sybase

## Technologies
- Hadoop
- Hadoop
- Mahout
- HBase
- Cassandra

---

Copyright © 2012 Dave Fainleib
dave@vcf/dave.com
blogs.forbes.com/davefainleib

Distributed stream processing tools

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

• Storm, S4, Hadoop 2.0 YARN, StratoSphere (.eu), BSP: Hama, Giraph, ... ?
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/](http://storm-project.net/)
A framework for real-time prediction
A framework for real-time prediction

- User history
- Cell tower statistics collector
- Data aggregator
- User predictor

- User defined functions

- Abstract processing component
- Reporter
- Input data parser

- Mobile data processing framework

- Java object serialization
- Kryo

- Serialization

- Storm
- Zookeeper
- ZeroMQ

- Streaming

- Persistent key-value store

- Caching

- Kryo

- Persistence adapter

- Disk
- Cassandra
- HBase
- JDBC

- Persistence
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
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