Using Apache Flink for Smart Cities: Warsaw case study

4th September 2018

Piotr Wawrzyniak,
Jarosław Legierski, PhD

Orange Polska
Orange Labs R&D Centre

Agenda

1. VaVeL project
2. Vehicles Movement Analyser
3. Vehicle delay prediction
4. Cluster deployment and integration
5. Operational management
6. Summary

VaVeL: Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors

This research has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688380 'VaVeL: Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors'.

VaVeL

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

Technion

IBM

orange

Fraunhofer

AGT international

Dublin City Council

CITY OF WARSAW

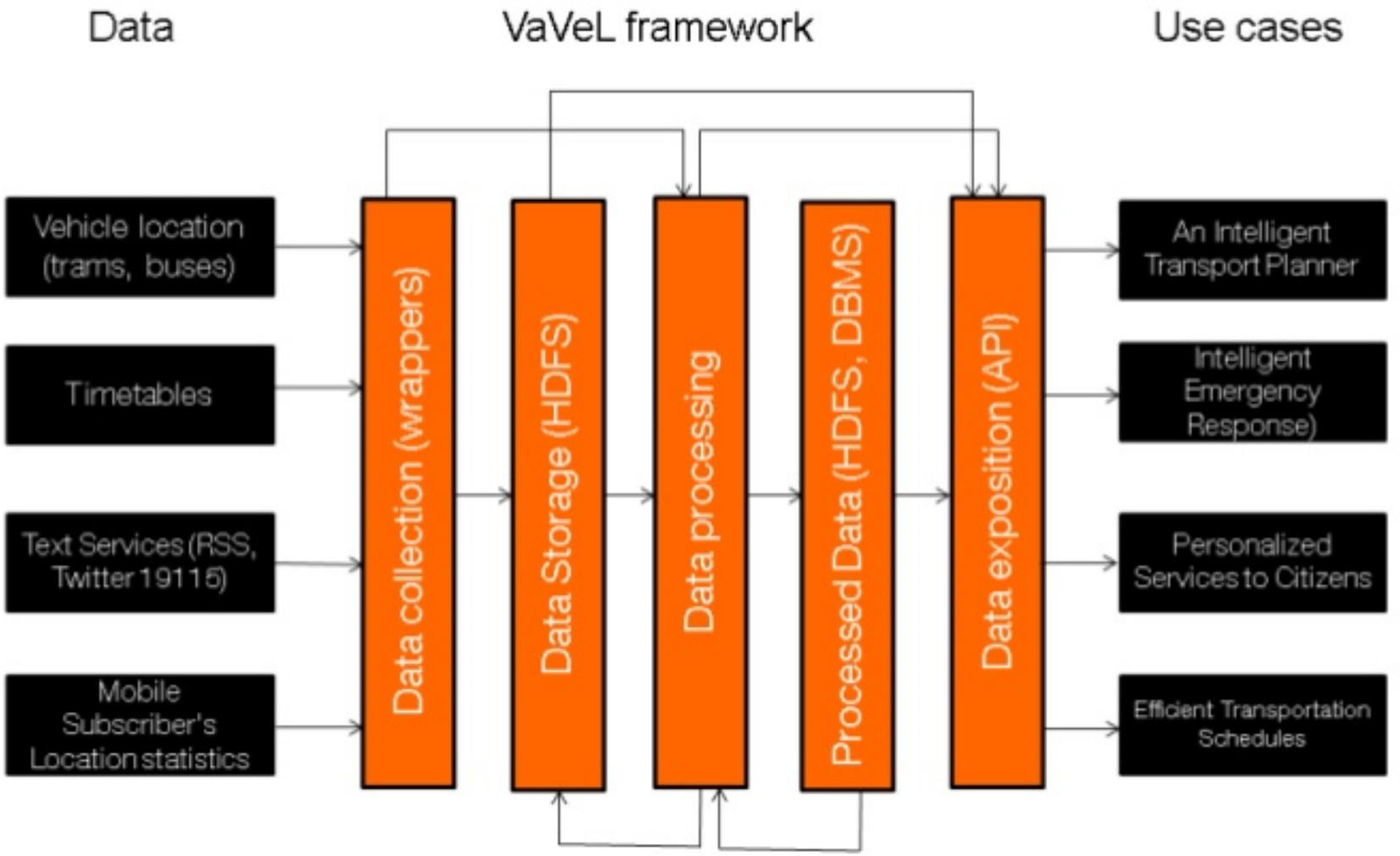
VaVeL project

1. The goal of the VaVeL project is to radically advance our ability to use urban data in applications that can identify and address citizen needs and improve urban life. Our motivation comes from problems in urban transportation.
2. This project will develop a general purpose framework for managing and mining multiple heterogeneous urban data streams for cities become more efficient, productive and resilient. The framework will be able to solve major issues that arise with urban transportation related data and are currently not dealt by existing stream management technologies.
3. VaVeL aims at making fundamental advances in addressing the most critical inefficiencies of current (big) data management and stream frameworks to cope with emerging urban sensor data thus making European urban data more accessible and easy to use and enhancing European industries that use big data management and analytics.

City of Warsaw use case

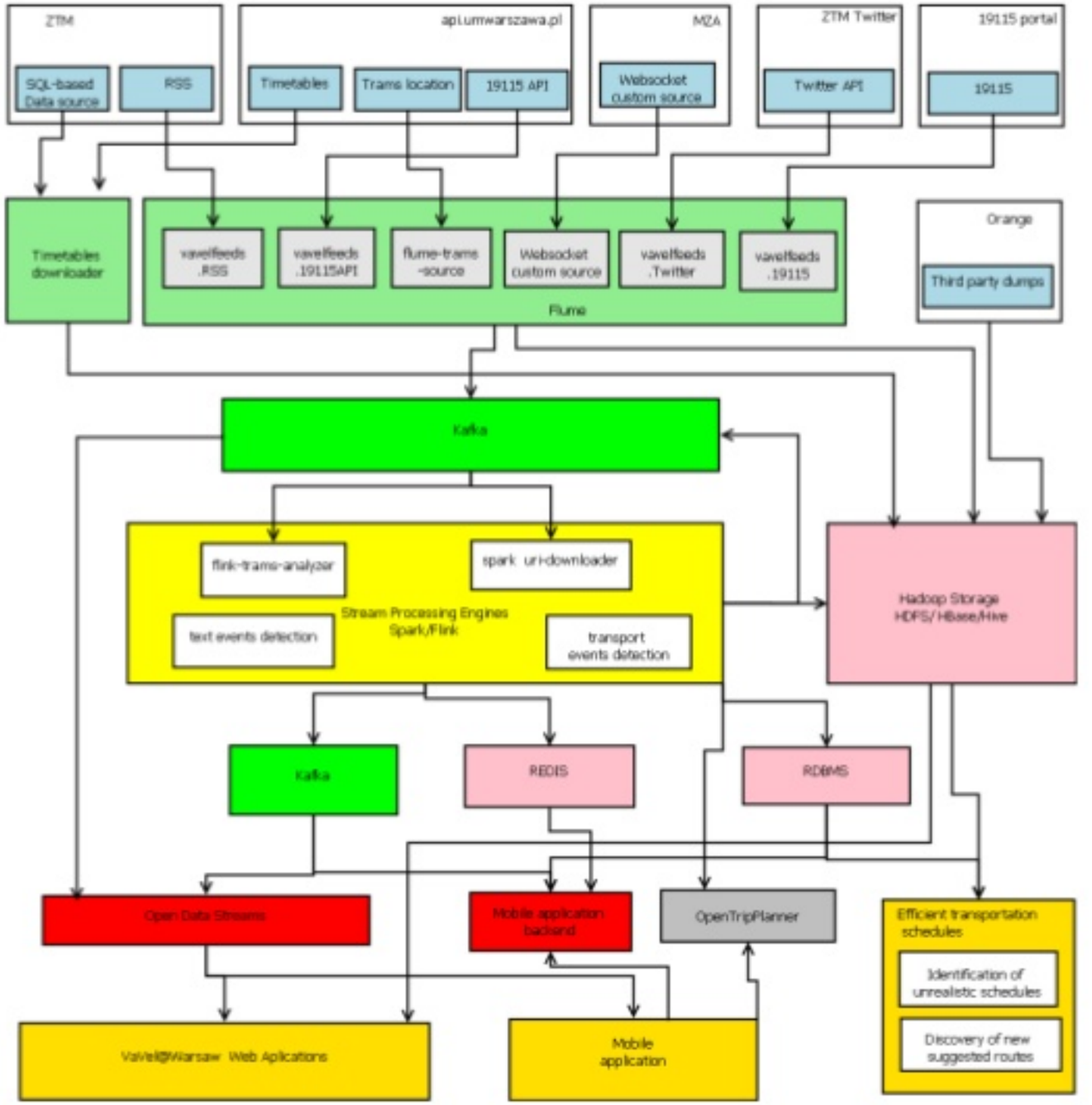
1. Warsaw use case focuses primarily on the public transport data
2. Two simultaneous streams of data:
   - realtime locations of trams (400+ vehicles during peak hours)
   - realtime locations of buses (1800+ vehicles during peak hours)
3. Moreover, static data are used to enrich raw data streams (timetables, maps etc.)
4. This is a joint work of Orange Polska and Warsaw University of Technology teams, in particular of Marcin Luckner, PhD, Karolina Kwasiborska and Tomasz Zaremba

City of Warsaw use case

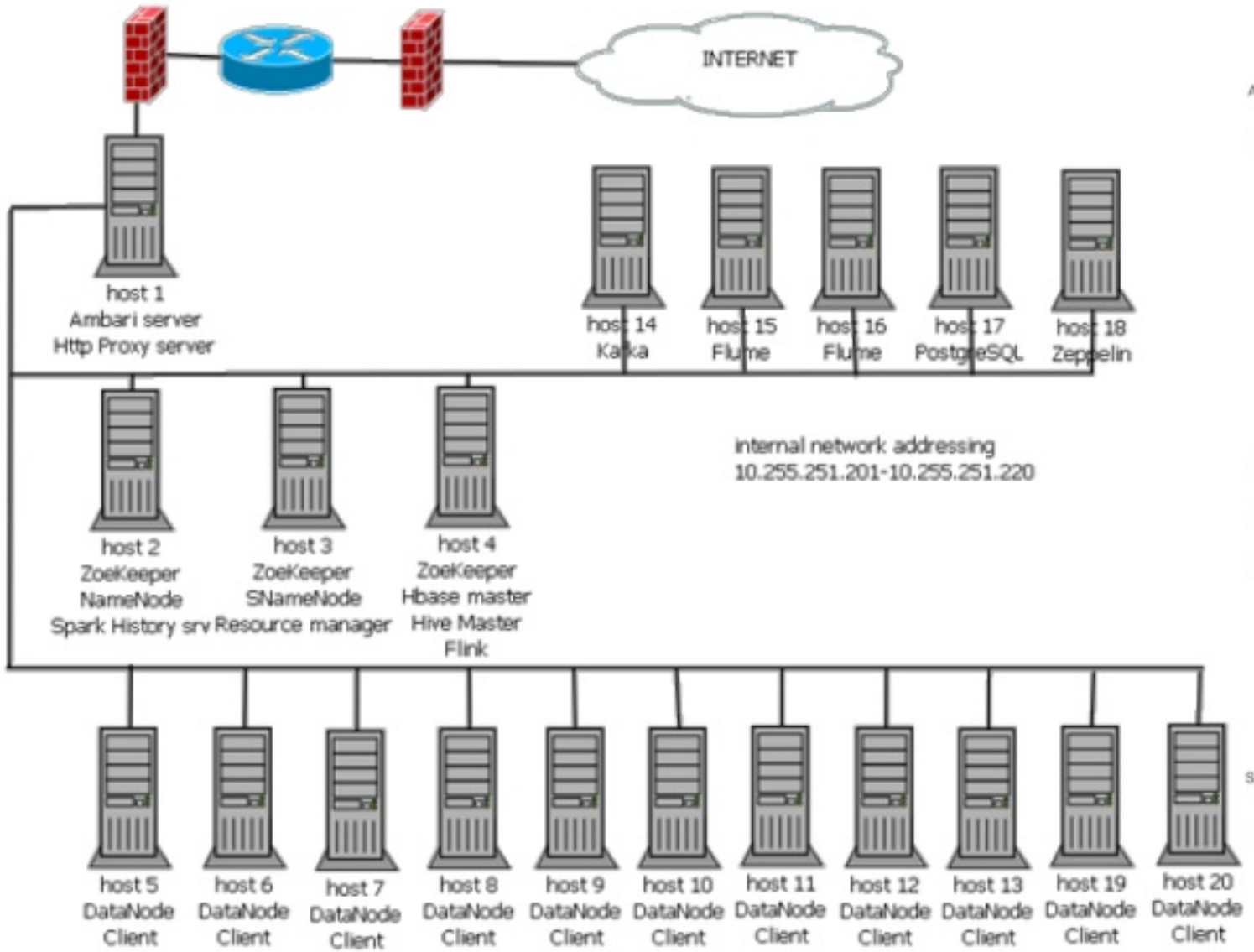


VaVeL system architecture

- Data sources layer
- Data acquisition layer - Apache Flume or dedicated applications
- Data processing layer – Apache Flink and Apache Spark
- Data Storage layer – HDFS, Redis, PostgreSQL
- Data Exposition layer – NGNIX, Open Trip Planner
- Application layer – mobile application, web applications



Hadoop cluster



Assign Masters

Assign master components to hosts you want to run them on.

NameNode:	host2.bajorko.com (11.6 GB, 4 c
SNameNode:	host3.bajorko.com (11.6 GB, 8 c
App Timeline Server:	host3.bajorko.com (11.6 GB, 8 c
ResourceManager:	host3.bajorko.com (11.6 GB, 8 c
History Server:	host3.bajorko.com (11.6 GB, 8 c
Hive Metastore:	host4.bajorko.com (11.6 GB, 4 c
WebHCat Server:	host4.bajorko.com
HiveServer2:	host4.bajorko.com (11.6 GB, 4 c
HBase Master:	host4.bajorko.com (11.6 GB, 4 c
ZooKeeper Server:	host2.bajorko.com (11.6 GB, 4 c
ZooKeeper Server:	host3.bajorko.com (11.6 GB, 8 c
ZooKeeper Server:	host4.bajorko.com (11.6 GB, 4 c
Metrics Collector:	host1.bajorko.com (11.6 GB, 4 c
Grafana:	host1.bajorko.com (11.6 GB, 4 c
Kafka Broker:	host14.bajorko.com (11.6 GB, 4
Spark History Server:	host2.bajorko.com (11.6 GB, 4 c
Zeppelin Notebook:	host18.bajorko.com (11.6 GB, 4
Flink:	host3.bajorko.com (11.6 GB, 8

host1.bajorko.com (11.6 GB, 4 cores)
- Metrics Collector
- Grafana

host14.bajorko.com (11.6 GB, 4 cores)
- Kafka Broker

host18.bajorko.com (11.6 GB, 4 cores)
- Zeppelin Notebook

host2.bajorko.com (11.6 GB, 4 cores)
- NameNode
- ZooKeeper Server
- Spark History Server

host3.bajorko.com (11.6 GB, 8 cores)
- SNameNode
- App Timeline Server
- ResourceManager
- History Server
- ZooKeeper Server
- Flink

host4.bajorko.com (11.6 GB, 4 cores)
- Hive Metastore
- WebHCat Server
- HiveServer2
- HBase Master
- ZooKeeper Server

Testbed OPL

Apache Flink – vehicles movement analyser

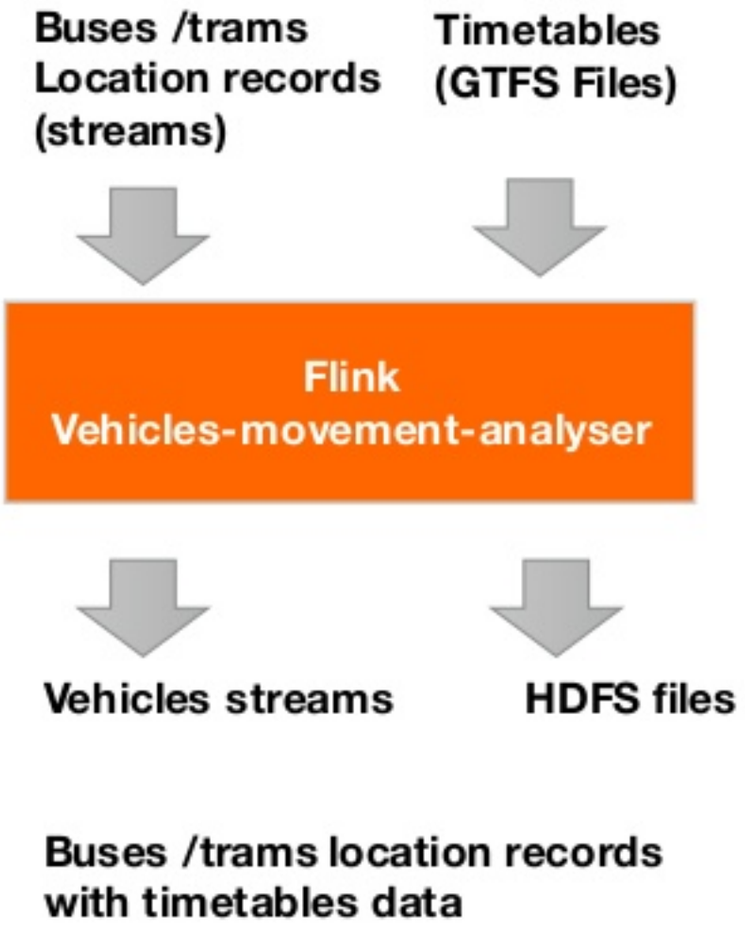
- Application combining vehicle location stream with static timetables data
- Computes delays, stops passing, position with respect to the schedule etc.
- Input:

Vehicles positions (Apache Kafka stream)

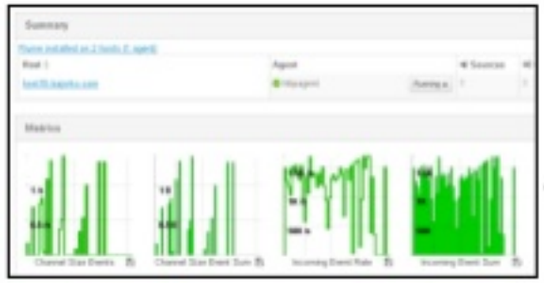
Timetables - GTFS files (General Transit Feed Specification)

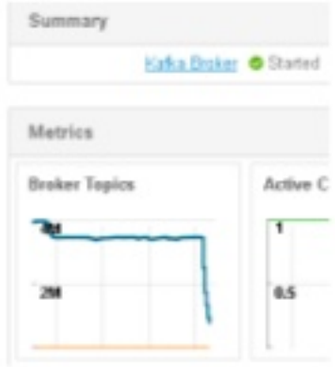
- Output:

Vehicles positions with timetables information (Apache Kafka stream & HDFS files)



Data Flow





api.um.warszawa.pl

Kafka

Flume
trams-source
buses-source

Timetables
downloader

HDFS

Flink
Vehicles-movement-
analyser

application_1478525427375_0031 flink Flink Application: pl.edu.pw.mini.streaming.TramInfoStreamCount Apache Flink default Fri Nov 18 04:00:21 +0100 Fri Nov 18 04:00:43 +0100 2016 FINISHED SUCCE

Input data

..

127;6;2018-08-29 09:27:25;20.89547;52.208771;2018-08-29T09:27:36.458
217;3;2018-08-29 09:27:23;21.024117;52.179916;2018-08-29T09:27:36.458
503;9;2018-08-29 09:27:25;21.053551;52.14637;2018-08-29T09:27:36.458
503;11;2018-08-29 09:27:18;21.043854;52.174828;2018-08-29T09:27:36.458
503;12;2018-08-29 09:27:26;21.024748;52.220844;2018-08-29T09:27:36.458
503;1;2018-08-29 09:27:26;21.017859;52.23724;2018-08-29T09:27:36.458
503;3;2018-08-29 09:27:29;21.017715;52.237766;2018-08-29T09:27:36.458
127;1;2018-08-29 09:27:28;21.019469;52.246325;2018-08-29T09:27:36.458
503;5;2018-08-29 09:27:25;21.042835;52.194279;2018-08-29T09:27:36.458
503;7;2018-08-29 09:27:21;21.042475;52.14782;2018-08-29T09:27:36.458
708;1;2018-08-29 09:27:30;20.795759;52.297321;2018-08-29T09:27:36.458
503;6;2018-08-29 09:27:24;21.034723;52.165955;2018-08-29T09:27:36.458
217;2;2018-08-29 09:27:28;21.07188;52.160404;2018-08-29T09:27:36.458

..

line String, brigade INT, time TIMESTAMP,lon DOUBLE, lat DOUBLE, finaltime TIMESTAMP

Output data

..

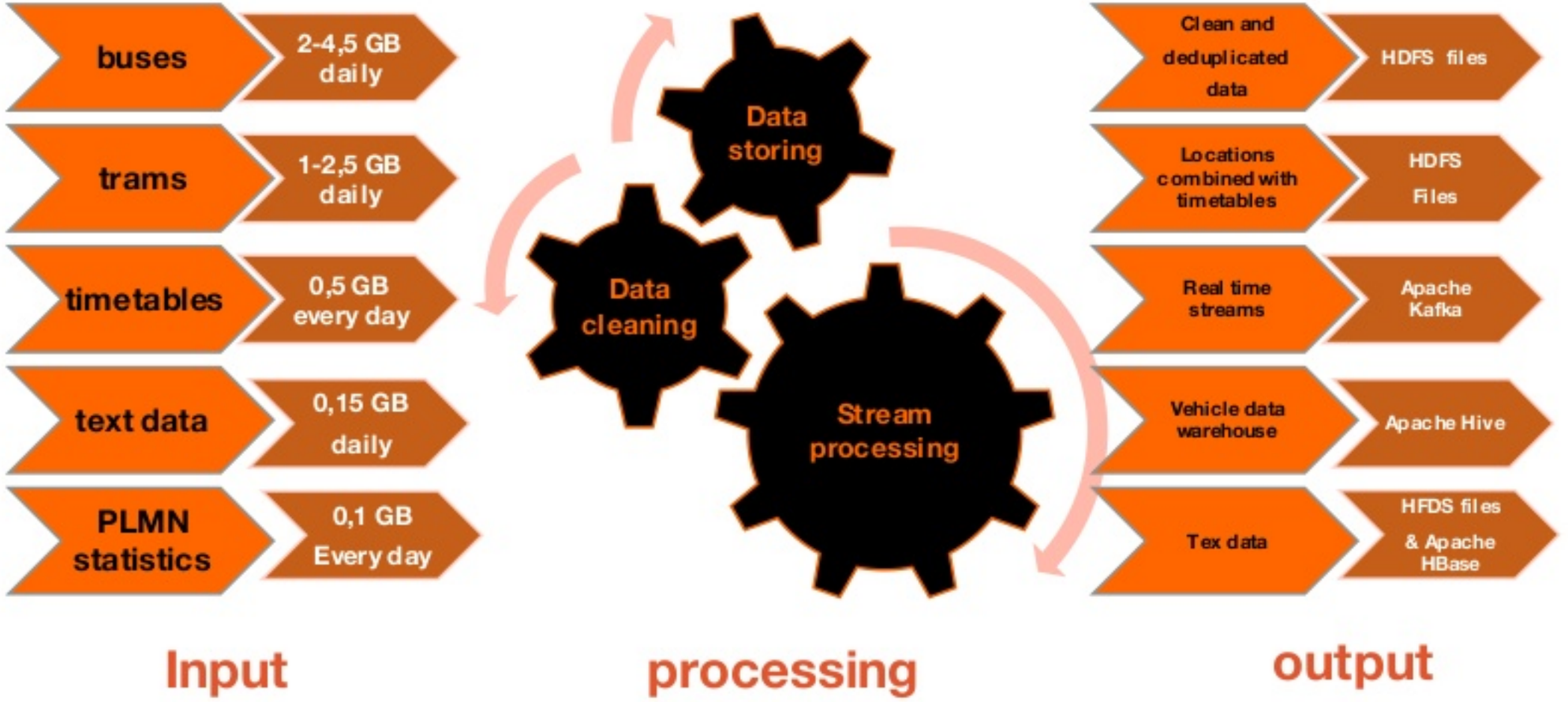
20180330;217;1;2018-08-29 09:29:15;21.041969;52.175838;21.041969;52.175838;MOVING;3.0;3038-Dominikańska;2018-08-29 09:30:00;3039-Dolina Służewiecka;72.7427662616879 3;21.043;52.17567;3039-Dolina Służewiecka;21.043;52.17567;72.7427662616879 3;2018-08-29 09:29:03;2018-08-29 09:29:03;3038-Dominikańska;21.03359;52.17739;596.841480655052 8;2018-08-29 09:30:00;217_1_113_0919;Metro Wilanowska;2018-08-29 09:19:00-2018-08-29 09:32:00;UNSAFE;2018-08-29T09:29:27.883;2018-08-29T09:29:28.532;false;false;;false;5.01;3.0;OPL;10;9;10;3038_02;3039_04;3038_02;3009_17

..

Output data

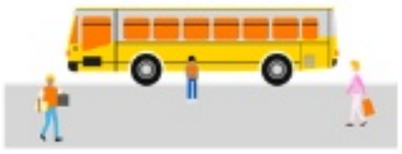
version STRING, line String, brigade INT, time TIMESTAMP,lon DOUBLE, lat DOUBLE, rawLon DOUBLE, rawLat DOUBLE, status STRING, delay STRING, delayAtStop STRING, plannedLeaveTime TIMESTAMP, nearestStop STRING, nearestStopDistance DOUBLE, nearestStopLon DOUBLE, nearestStopLat DOUBLE, previousStop STRING, previousStopLon DOUBLE,previousStopLat DOUBLE, previousStopDistance DOUBLE, previousStopArrivalTime TIMESTAMP, previousStopLeaveTime TIMESTAMP, nextStop STRING, nextStopLon DOUBLE, nextStopLat DOUBLE, nextStopDistance DOUBLE, nextStopTimetableVisitTime TIMESTAMP, courseIdentifier STRING, courseDirection STRING, timetableIdentifier STRING, timetableStatus STRING, receivedTime TIMESTAMP, processingFinishedTime TIMESTAMP, onWayToDepot BOOLEAN, overlapsWithNextBrigade BOOLEAN, atStop STRING,overlapsWithNextBrigadeStopLineBrigade STRING, speed DOUBLE, (serverId STRING) delayAtStopStopSequence DOUBLE ,previousStopStopSequence DOUBLE,nextStopStopSequence DOUBLE,delayAtStopStopId STRING,previousStopStopId STRING,nextStopStopId STRING

Data volume


buses
2-4,5 GB daily
trams
1-2,5 GB daily
timetables
0,5 GB every day
text data
0,15 GB daily
PLMN statistics
0,1 GB Every day
Data storing
Data cleaning
Stream processing
Clean and deduplicated data
HDFS files
Locations combined with timetables
HDFS Files
Real time streams
Apache Kafka
Vehicle data warehouse
Apache Hive
Tex data
HFDS files & Apache HBase
Input
processing
output

Delay prediction module- motivation

Module Architecture



Data input

Data Preprocessing Module

Trams

5 minutes window

10 minutes window

n minutes window

Buses

5 minutes window

10 minutes window

n minutes window

Testing Streams

Learning Streams

Machine Learning Module

Delay prediction

5 minutes period

10 minutes period

n minutes period

Position prediction

5 minutes period

10 minutes period

n minutes period

Apache SAMOA

Streams of predictions

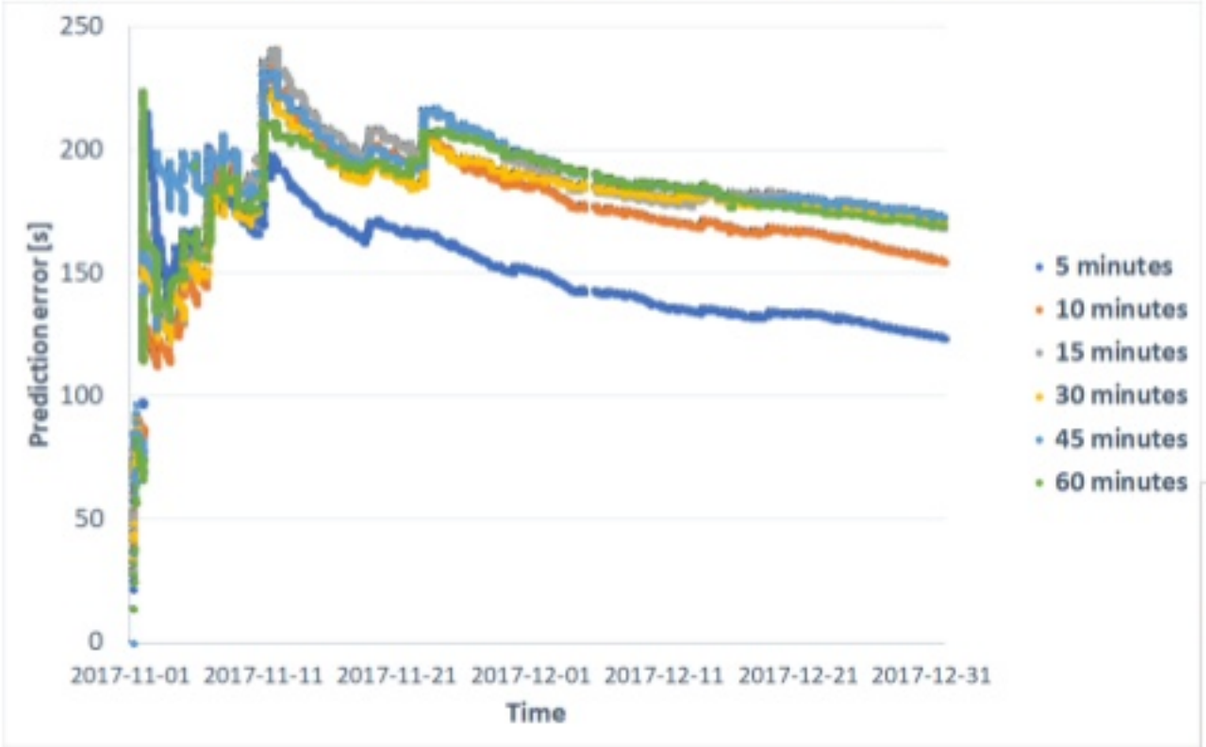
Data Postprocessing Module

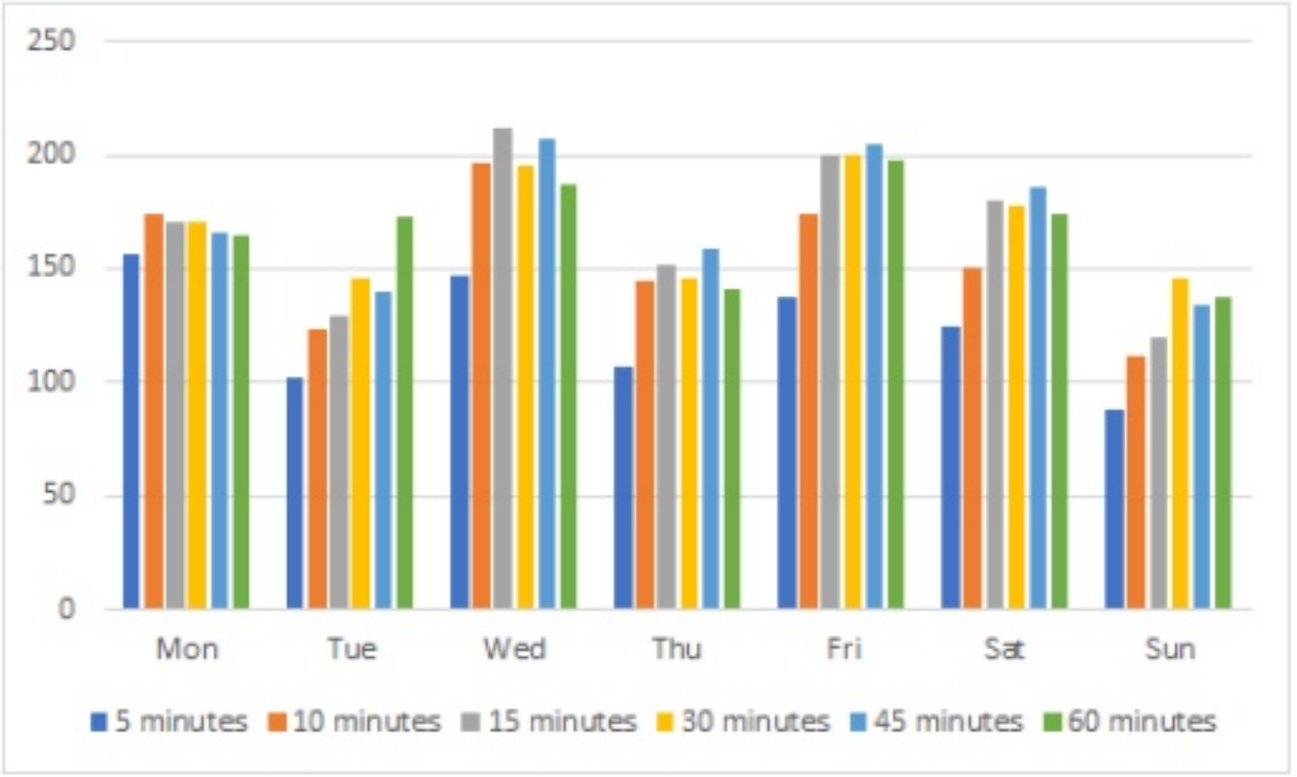
General-purpose formating

GTFS

Additional data merger

ML methods used for validation





Evaluation results

VAMR model of Apache SAMOA

Prediction period	Delay [s] mean	Delay [s] stdev	Position [m] mean	Position [m] stdev
5	123.24	400.26	37.98	25.1
10	153.43	496.2	38.79	25.05
15	166.15	524.53	41.57	25.42
30	168.75	533.73	43.94	24.3
45	170.91	598	44.27	24.28
60	168.13	492.48	47.7	26.13

MLP models of WEKA

Prediction period	Delay [s] mean	Delay [s] stdev	Position [m] mean	Position [m] stdev
5 minutes	133.7	294.34	45.2	611.35
10 minutes	141.32	336.06	40.07	687.75
15 minutes	135.83	348.93	183.77	298.93
30 minutes	139.86	329.72	66.16	587.6
45 minutes	138.54	346.4	56.44	224.7
60 minutes	140.56	317.61	62.63	178.74

Operational management

1. Cluster (incl. Apache Flink) is manager via Apache Ambari
2. Important: Ambari creates empty Flink job to monitor the cluster, so you should see n+1 jobs, where n is the number of ‚real' Flink applications.

Thank you ☺

This research has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688380 'VaVeL: Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors'.

visit www.vavel-project.eu for info on the VaVeL project

VaVeL

orange™

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

tu

Technion

IBM.

Fraunhofer

AGT international

Dublin City Council
Comhairle Cathrach Bhaile Átha Cliath

CITY OF WARSAW