Grant Agreement No: 687591

14/7/17

Big Data Analytics for Time Critical Mobility Forecasting

datAcron

D1.7 Integrated prototype (interim)

Deliverable Form	
Project Reference No.	H2020-ICT-2015 687591
Deliverable No.	1.7
Relevant Work Package:	WP 1
Nature:	R
Dissemination Level:	CO
Document version:	1.0
Due Date:	30/06/2017
Date of latest revision:	14/07/2017
Completion Date:	14/07/2017
Lead partner:	UPRC
Authors:	Christos Doulkeridis (UPRC), Apostolos Glenis (UPRC), Panagiotis Nikitopoulos (UPRC), Giorgos Santipantakis (UPRC), Yannis Theodoridis (UPRC), Akrivi Vlachou (UPRC), George Vouros (UPRC), Michael Mock (FRHF), Nikos Pelekis (UPRC), Harris Georgiou (UPRC), Kostas Patroumpas (UPRC), Elias Alevizos (NCSR'D), Georg Fuchs (FRHF)
Reviewers:	Yannis Theodoridis (UPRC), Michael Mock (FHRF)
Document description:	This deliverable briefly describes the datAcron integrated prototype.
Document location:	WP1/Deliverables/D1.7/Final

HISTORY OF CHANGES

Version	Date	Changes	Author	Remarks
0.1	15/6/2017	First version of report	C. Doulkeridis	
0.9	12/7/2017	Near-final version	C.Doulkeridis, A.Glenis, P.Nikitopoulos, G.Santipantakis, A.Vlachou, G.Vouros	Version shared with all partners
1.0	14/7/2017	Final version	C.Doulkeridis, A.Glenis, P.Nikitopoulos, G.Santipantakis, A.Vlachou, G.Vouros	

EXECUTIVE SUMMARY

This report is the seventh deliverable (D1.7) of datAcron work package 1 "System architecture and data management" with main objective to describe the datAcron integrated prototype system, in accordance with the requirements specified in deliverable D1.1, the architecture specified in deliverable D1.2, and the software design described in D1.6.

Essentially, D1.7 is a brief, summarized version of D1.6, mainly describing the integrated system components, the flows and functionality provided, focusing on those aspects of the integrated prototype intended as demonstration in the context of the project.

TABLE OF CONTENTS

TERMS & ABBREVIATIONS

ADS-B	Automatic Dependent Surveillance-Broadcast
AIS	Automatic Identification System
RDF	Resource Description Framework

LIST OF FIGURES

LIST OF TABLES

1 Introduction

This document is the deliverable D1.7 “Integrated prototype (interim)” of Task T1.3 of work package 1 ”System Architecture and Data Management” of the datAcron project. It provides a brief description of the datAcron integrated prototype and its status as of M18, as it is planned to be demonstrated in the first review of the project. Recall that the integrated prototype is the result of the first iteration of software design, aiming at providing an early prototype and integrating software modules towards the second iteration. The second design iteration focuses on the integration with WP2–WP4.

1.1 Purpose and Scope

The “Integrated prototype (interim)” aims at demonstrating a fast prototype which encompasses the basic, integrated functionality of datAcron. In particular, the prototype that is developed on M18 provides integrated functionality, by implementing interconnection between modules developed as results of different work packages. In more detail, the prototype offers the following functionality on M18:

- Real-time consumption of raw surveillance data streams, on which the following operations are performed: in-situ processing (including low-level event detection and synopsis generation), semantic integration with weather and contextual data sources, future location prediction, complex event detection and forecasting, and real-time visualization.
- Querying integrated RDF data with a limited set of spatio-temporal queries, demonstrating the feasibility of query processing.

1.2 Approach for the Work package and Relation to other Deliverables

Similarly to D1.6, D1.7 is based on D1.2 “Architecture Specification” delivered in M12, as it builds on and refines D1.2 into a more concrete software architecture, taking also into account the features offered by the technological solutions adopted. It should be pointed out that D1.7 is prepared during the same time that deliverables D1.3 “Cross-streaming data processing (interim)”, D1.4 “Data Integration, Management (interim)”, D2.1 “’Cross-streaming, real-time detection of moving object trajectories (interim)’, D1.5 “Data storage and querying (interim)”, and D1.6 “Software design (interim)” are prepared. Deliverables D1.3, D1.4 and D2.1 provide detailed description on real-time layer of the datAcron architecture. Also, D1.5 describes the design of the batch processing solution developed in the context of datAcron. As such, they connect to this deliverable, and intermediate versions of D1.3, D2.1 and D1.5 have been considered during the preparation of the current deliverable. Most importantly, D1.6 essentially defines the software design of the integrated prototype.

1.3 Methodology and Structure of the Deliverable

In terms of work methodology, this deliverable takes as input D1.6 "Software design (interim)" and develops a prototype system according to the specification of D1.6. The prototype has been developed in close collaboration with partners in WP2, WP3, and WP4, that develop their solutions and prototype implementations, in order to provide specific guidelines on the technology used for inter-communication of modules, and for coordinating and orchestrating the development activities.

The remaining of this report is structured as follows:

- Section 2 presents the software design of the integrated prototype of datAcron, focusing in flows of information and the architecture that encompasses all envisioned flows and modules. Moreover, it reports on the implementation status of the integrated prototype and provides details on indicative, yet critical, implemented information flows.
- Section 3 summarizes the work reported in this deliverable, and provides the plan for the second iteration.

2 The datAcron Integrated Prototype

The purpose of this section is to provide a description of the integrated prototype of datAcron and clarify the details of each module / flow as of M18. Therefore, in Section 2.1, we present details of participating modules, and, in Section 2.2, we discuss the respective flows materialized by these modules. Section 2.3 summarizes the results presented regarding the datAcron integrated system prototype.

2.1 datAcron Modules

The envisaged datAcron architecture consists of the following modules listed in Table 1.

Module Nr.	Module acronym	Module Title	Task in Charge	Work package
1)	Maritime	Maritime raw data feeder	T5.2	WP5
2)	Aviation	Aviation raw data feeders	T6.2	WP6
3)	LED	In-situ processing 1 – Low-level event detector	T1.3.1	WP1
4)	SG	In-situ processing 2 – Synopses generator	T2.1	WP2
5)	SI	Semantic integrator	T1.3.2	WP1
6)	DM	Data manager	T1.3.3	WP1
7)	T/FLP	Trajectory / Future location predictor	T2.2	WP2
8)	TDA	Trajectory data analytics	T2.3	WP2
9)	CER/F	Complex event recognition / forecasting (CER/F)	T3.1-2-3	WP3
10)	IVA	Interactive visual analytics	T4.3	WP4
11)	Viz	Real-time visualization	T4.4	WP4

Table 1: The modules in the datAcron integrated prototype.

A short description as well as the implementation status (as of M18) of each module follows.

- **Maritime raw data feeder (Maritime)**: The data feed is a decimated stream that comes from a range of terrestrial AIS receivers and 18 satellites in a low earth orbit. The maritime AIS data stream is collected, tested for veracity using a streaming analytics module and then filtered to provide the data required for the datAcron project. The AIS data stream is then converted, in real time, from an IEC 61162-1 data stream to a JSON format data stream to allow it to be ingested into the remainder of the datAcron system.
- **Aviation raw data feeder (Aviation)**: This module comprises of a set of 6 European Surveillance data feeds. Namely:
    1. FlightAware real time surveillance feed: This module sends a stream of real time surveillance from flightaware global live feed data in plain text to the stdout so it can be piped to any consumer. Only one connection is allowed for datAcron project.

2. FlightAware replay surveillance feed, online mode: This module sends a stream of data for a given period in the past. It requires internet connection and uses flight aware services. The data streamed starts at the beginning of the period and last till all messages are delivered to the end of the period (a few days are maximum period span allowed).
3. FlightAware replay surveillance feed, offline mode: This module sends a stream of data reading a local json file previously stored from the real time surveillance feed. It does NOT require internet.
4. ADSBExhcange real time surveillance feed: This module sends a stream of real time surveillance from ADSBExchange global live feed data in plain text to the stdout so it can be piped to any consumer.
5. ADSBExhcange replay surveillance feed, offline mode: This module sends a stream of data reading a local file previously stored from the real time ADSBExchange surveillance feed. It does NOT require internet.
6. ADSBHub replay surveillance feed, offline mode: This module sends a stream of data reading a local file previously stored from the real time ADSBHub surveillance feed. It does NOT require internet.

- **In-situ processing 1 – Low-level event detector (LED)**: In-situ processing in general refers to the ability to process data streams as close to the source where the data originates. This is in particular challenging, when the stream processing of the data requires additional input from other sources, either other instances of the stream or from global system settings or user interaction. As a proof of achieving the architectural integration of in-situ processing, datAcron will apply the event forecasting techniques in-situ on single streams, providing an enriched event stream for visualization in real-time. Based on this, event forecasting will be extended towards distributed online learning, making it possible to learn forecasting models cross-stream from other moving objects.

- **In-situ processing 2 - Synopses Generator (SG)**: The Synopses Generator consumes streaming positions of raw surveillance data and eliminates any inherent noise such as delayed or duplicate messages. Moreover, it identifies critical points along each trajectory, such as stop, turn, or speed change, in order to provide an approximate, lightweight synopsis per moving object.

- **Semantic integrator (SI)**: Its functionality is to (a) transform data from all sources to RDF and (b) discover links between different sources, output this as a stream and also send for storage.

- **Data manager (DM)**: Its functionality is to store information into a distributed spatio-temporal RDF store and provide query answering facility.

- **Trajectory / Future location predictor (T/FLP)**: FLP calculates motion functions by harvesting the cleansed KAFKA stream (from the Synopses Generator module) consisting of the most recent locations from a moving object to predict its short-term future location in real time by taking into consideration the tendency of the movement. Each predicted point will be streamed out in real time to other modules. Regarding TP, it will present a similar functionality targeting at predicting the future trajectory of a moving object as far in time horizon as possible.

- **Trajectory data analytics (TDA)**: The goal of this module is twofold: on the one hand it provides advanced analytics that are going to serve specialized requirements in the datAcron architecture (e.g. data-driven discovery of the networks/routes upon which the movement of the vessels/aircrafts take place), while on the other hand it provides global patterns that represent meta models devised from the local patterns (e.g. clusters and sequential patterns of semantic trajectories).
- **Complex event recognition/forecasting (CER/F)**: CER is about real-time detection of complex events, whereas CEF is about real-time forecasting of complex events. Both modules are working on the synopsis of the moving object generated by the two in-situ processing modules (LED & SG) and, in addition, can take additional information achieved via the enrichment and linking performed by SI. The output of CER is a real-time stream with detected events. On the other hand, CEF enriches the input stream with a forecast about the probability of each monitored pattern. As event forecasting requires learning of CER probabilities, it is more restrictive with respect to the potentially supported patterns than the pure event detection.
- **Interactive visual analytics (IVA)**: The IVA module builds on top of the real-time visualization module to provide limited analytical capacity on streaming data. The primary use is to allow analysts, and possibly advanced operators, to fine-tune and observe impact of parameter adjustments to the T/FLP and CER/F modules compared to actual data in (near) real-time. It therefore complements the situation monitoring capabilities of the real-time visualization used by ordinary operators on the one hand (by providing parameter settings to the detection modules), and the full-fledged VA suite used for in-depth exploration and analysis in offline (strategic latency) settings.
- **Real-time visualization (Viz)**: This module provides a map-based visualization of the stream of enriched spatio-temporal events generated by the T/FLP and CER/F modules. It is able to display different event types (e.g., critical points) simultaneously with individual visual encoding for each type. In addition events associated with the same moving object identifier are automatically integrated into trajectory representations so operators can observe movement patterns. The overall design follows the "overview-first, zoom-and-filter, details-on-demand" approach, meaning that operators can define filters on the input stream to drill down on areas and event types of interest.

2.2 datAcron Flows

Having the modules presented above in hands, a number of information flows of three different types are envisaged. In particular:

- *Information management* flows are about the reconstruction of trajectories and their enrichment with useful annotation, which is to be performed online (operational latency), and their storage for querying purposes, which is to be performed offline (tactical latency).
- *Online analytics* flows are about consuming the available streaming information, which is to be performed online (operational latency); and
- *Offline analytics* flows are about consuming the available stored information, which is to be performed offline (strategic latency).

Note that there exist three main consumers (namely, T/FLP, CER/F, and IVA), therefore, 3+3 flows are envisaged, for the online and offline analytics, respectively. Table 2 presents the list of flows (along with the respective latency type and partners in charge of coordinating their implementation).

Flow Nr.	Flow Title	Latency
Information management flows		
1)	Trajectory reconstruction and semantic enrichment	Operational
2)	RDF storage	Tactical
Online analytics flows		
3)	Trajectory/FL prediction online	Operational
4)	Complex event recognition / forecasting online	Operational
5)	Visual Analytics online	Tactical
Offline analytics flows		
6)	Trajectory data analytics offline (*)	Strategic
7)	Complex event recognition / forecasting offline (*)	Strategic
8)	Visual Analytics offline	Strategic

Table 2: The datAcron flows of information (note: flows marked with * are not planned to be implemented until M18).

The functionality of each flow is discussed in the following sections. In accordance with the flows, Figure 1 illustrates the datAcron architecture, which is a refined version of the architecture specified in Deliverable D1.2 "Architecture Specification".

2.2.1 Flow #1: Trajectory reconstruction & semantic enrichment (operational latency)

- Short description: Maritime / Aviation raw data stream is (1a) cleansed, enriched with derived information (e.g. speed) as well as low-level events (e.g. intersection with zones of interest), synopsized by tagging "critical points" (change of heading or altitude, etc.), and (1b) further enriched with info from other external data sources / streams (weather info, etc.); the final output (1c) is streamed out to be consumed by other modules, including its visualization (1d).
- Modules involved: Maritime / Aviation; LED; SG; SI; Viz.

2.2.2 Flow #2: RDF storage (tactical latency)

- Short description: A subset of the enhanced surveillance data stream, i.e. the annotated surveillance data, as well as selected output streamed out by other modules is (2a) processed and (2b) stored in the RDF store.
- Modules involved: DM.

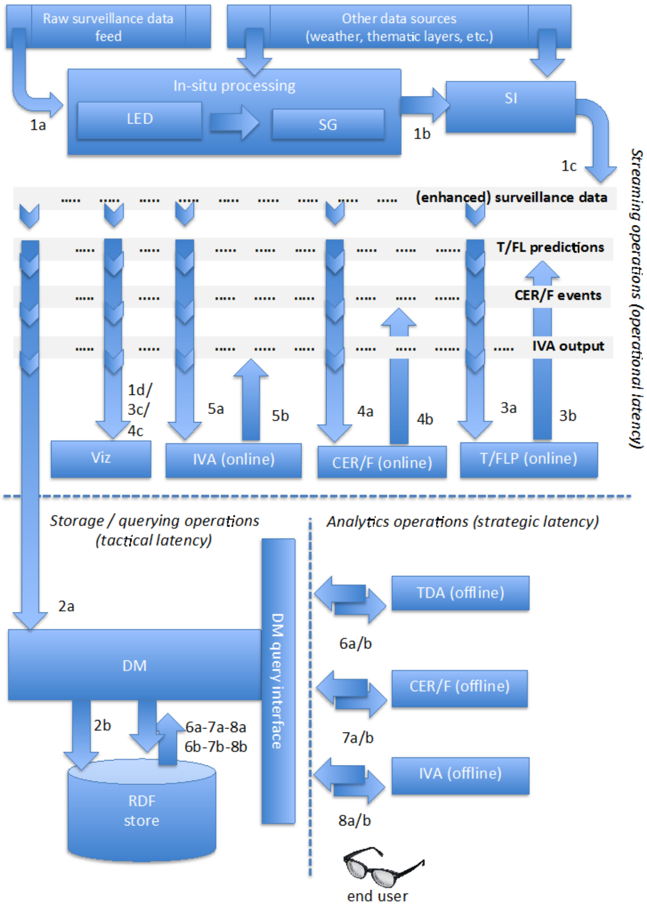

Figure 1: The refined datAcron architecture.

2.2.3 Flow #3: Trajectory/FL prediction online (operational latency)

- Short description: T/FLP (3a) consumes the enhanced surveillance data stream (as well as other streams, if needed) for the purposes of online trajectory / future location prediction and (3b) streams out its output to be consumed by other modules, including its visualization (3c).
- Modules involved: T/FLP; Viz.

2.2.4 Flow #4: Complex event recognition / forecasting online (operational latency)

- Short description: CER/F (4a) consumes the enhanced surveillance data stream (as well as other streams, if needed) for the purposes of online event recognition / forecasting and (4b) streams out its output to be consumed by other modules, including its visualization (4c).

- Modules involved: CER/F; Viz.

2.2.5 Flow #5: Interactive visual analytics online (operational latency)

- Short description: IVA consumes the enhanced surveillance data streams (3c, 4c), streamed meta data on the T/FLP and CER/F modules (current parameter settings, 5a), and, if needed, base data for comparison (1d) for the purposes of online VA; and (5b) streams out its output (updated parameter settings, areas-of-interest) in KVP format to be consumed by other modules.
- Modules involved: IVA; T/FLP; CER/F.

2.2.6 Flow #6: Trajectory data analytics offline (strategic latency)

- Short description: TDA (6a) queries the RDF store in order for complex patterns to be discovered and (6b) stores selected results back to the RDF store for future use.
- Modules involved: TDA; DM.

2.2.7 Flow #7: Complex event recognition / forecasting offline (strategic latency)

- Short description: CER/F (7a) queries the RDF store in order for complex events to be detected/forecasted and (7b) stores selected results back to the RDF store for future use.
- Modules involved: CER/F; DM.

2.2.8 Flow #8: Interactive visual analytics offline (strategic latency)

- Short description: IVA (8a) queries the RDF store in order for large batches of raw data for complex offline analysis and (8b) stores selected results (derived attributes, spatio-temporal patterns, clustering results, parameter settings) back to the RDF store for future use.
- Modules involved: IVA; DM.

2.3 Summary

This section presented datAcron modules and the design of respective flows involving these modules, thereby demonstrating the integrated prototype and its main functionality. All datAcron technical work packages provide software modules in an integrated prototype, and the prototype will be demonstrated in the first review of datAcron.

3 Summary & Outlook

This report accompanies the demonstration of the integrated prototype (D1.7) and provides a short description of it. It presents different flows of information in datAcron, the modules that comprise the integrated prototype, their intercommunication, and the implementation status as of M18.

In the second iteration, the plan is to improve this prototype and extend its functionality in different ways:

- extend the functionality of individual modules, in order to optimize their internal operation,
- evaluate and if necessary refine the implementation to achieve the required latency constraints in real-time processing,
- provide a richer suite of queries over the distributed RDF store, thereby enabling retrieval of enriched data sets using a wide variety of filtering criteria, which can be used for data analysis tasks,
- develop and integrate with WP2–WP4 the offline analytics, described as Flows #6–#8.