Track & Know

Big Data for Mobility Tracking Knowledge Extraction in Urban Areas

D1.3 Interoperability Requirements

Document Summary Information

Grant Agreement No	780754	**Acronym**	TRACK & KNOW
Full Title	**Big Data for Mobility Tracking Knowledge Extraction in Urban Areas**		
Start Date	01/01/2018	**Duration**	36 months
Project URL	https://trackandknow.eu		
Deliverable	D1.3 Interoperability Requirements		
Work Package	WP1:		
Contractual due date	31.8.18	**Actual submission date**	31.8.18
Nature	Report	**Dissemination Level**	PU
Lead Beneficiary	ILS		
Responsible Author	Ibad Kureshi (ILS)		
Contributions from	Luk Nappen, Toni Staykova, Leonardo Longhi, Fabio Manichetti, Ian Smith, Mirco Nanni, Athanasios Koumparos, Anagnostis Delkos, Panos Livanos, Marios Logothetis, Ioannis Daskalopoulos, Gennady Andrienko, Ioannis Daskalopoulos		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 780754.

Executive Summary

This report outlines the Track&Know platform and pilots' compliance with the United States' National Institute of Standards and Technologies' Big Data Interoperability Framework. The framework provides a reference architecture against which not only different technical components but also applications and analytics can be mapped. Known as the NIST Big Data Reference Architecture (NBDRA), it provides a technology-, infrastructure-, and vendor-agnostic framework to ensure interoperability of applications, components, and systems.

The NBDRA method of compliance is through the mapping of activities along industrial verticals and data value chains. Therefore, future adopters of mobility analytics can map their analysis workflow to the NBDRA and in-turn to the various components developed as part of the Track&Know project.

This deliverable explains the salient features of the NBDRA, including the recommended standards, and then maps the proposed Track&Know pilots, toolboxes and platform to the framework. Mechanisms and safeguards to ensure compliance to the Interoperability ethos in an ever-evolving space are also described. This challenge of ever evolving compliance requirements are highlighted throughout the report, particularly as during the preparation of this report the NBDRA itself was significantly updated.

The European Commission too published its report on best practices for Interoperability during the preparation of this report. An overview of this report is included. The European approach to interoperability takes a different angle to the NIST framework and helps the Track&Know project fill some of the gaps in the NBDRA.

Revision history (including peer reviewing & quality control)

Version	Issue Date	% Complete	Changes	Contributor(s)
V0.1	26.3.18	5%	Initial Deliverable Structure	Ibad Kureshi (ILS)
V0.2	20.6.18	45%	NIST Details	Ibad Kureshi (ILS)
V0.3	31.6.18	70%	Pilot to NIST mapping	Ibad Kureshi, Luk Nappen, Toni Staykova, Leonardo Longhi, Fabio Manichetti, Ian Smith, Mirco Nanni, Athanasios Koumparos, Anagnostis Delkos, Panos Livanos, Marios Logothetis, Ioannis Daskalopoulos, Gennady Andrienko
V0.4	15.7.18	99%	Introduction, Conclusion, Infrastructure Mapping	Ibad Kureshi (ILS), Ioannis Daskalopoulos (INTRA)
V0.5	16.8.18	100%	Updating Framework, Alternate Frameworks	Ibad Kureshi (ILS)
V0.6	20.8.18	100%	Peer-Review Feedback	Cheng Fu (UZH), Athanasios Koumparos (ZEL)
V1.0	30.8.18	100%	QA Feedback	Toni Staykova (CEL)

Disclaimer

Copyright message

Table of Contents

1 Introduction 7
1.1 Mapping TRACK&KNOW Outputs 7
1.2 Deliverable Overview and Report Structure 8
2 Big Data Interoperability and Frameworks 9
2.1 NIST Big Data Interoperability Framework 9
2.2 Competing Frameworks 11
3 NIST Reference Architecture 12
3.1 Mapping Use Cases and Requirements 12
3.2 Conceptual Model and Infrastructural Layering 14
3.2.1 Big Data Application Provider 15
3.2.2 Big Data Framework Provider 15
3.3 Interconnections 16
3.3.1 High Level Principles 16
3.3.2 Component-Specific Interconnections 17
3.4 Standards 17
4 Technology Platform 18
4.1 Outline 18
4.2 Interoperability Mapping 19
4.3 Common Repository Schema 20
4.4 Ensuring Compliance 21
5 Pilot 1: Insurance 22
5.1 Outline 22
5.2 Interoperability Mapping 22
6 Pilot 2: Healthcare Service 24
6.1 Outline 24
6.2 Interoperability Mapping 24
7 Pilot 3: Fleet Management 26
7.1 Outline 26
7.2 Interoperability Mapping 26
8 Conclusions 28
Annex I: Ethics Proforma 29

List of Figures

Figure 3.1: NIST Big Data Reference Architecture 14
Figure 4.1: Track&Know High-Level Architecture 18

List of Tables

Table 1-1: Adherence to TRACK&KNOW's GA Deliverable & Tasks Descriptions 7
Table 3-1: Use Case Characterisation mapped to NBDRA 12
Table 3-2: NBDRA Component Definitions and Actors 13
Table 4-1: Component to NIST Architecture Layer Mapping 19
Table 5-1: Insurance Pilot: Business Question NBDRA Mapping 23
Table 6-1: Healthcare Service Pilot: Business Question NBDRA Mapping 25
Table 7-1: Fleet Management Pilot: Business Question NBDRA Mapping 27

Glossary of terms and abbreviations used

Abbreviation / Term	Description
API	Application Programmable Interface
BDA	Big Data Analytics
BDAP	Big Data Application Provider
BDP	Big Data Processing
BMDI	Big Mobility Data Integrator
BSP	Bulk Synchronous Parallel
CER	Complex Event Toolbox
CIO	Chief Information Officer
CSIRT	Computer Systems Incident Response Teams
DAG	Directed Access Graph
DCAT	Data Catalogue Vocabulary
EC	European Commission
ELT	Extract, Load, Transfer
EU/EUR	European Union
EV	Electric Vehicles
FMUC	Fleet Management Use Cases
FOSS	Free and Open Source Software
GA	Grant Agreement
GDPR	General Data Protection Regulation
GP	General Practitioner
HCUC	Healthcare Use Cases
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
INCITS	International Committee for Information Technology Standards
ISO	International Standards Organisation
IT	Information Technology
IUC	Insurance Use Cases
KPI	Key Performance Indicators

NBD-PWG	NIST Big Data Public Working Group
NBDIF	NIST Big Data Interoperability Framework
NBDRA	NIST Big Data Reference Architecture
NHS	National Health Service
NIST	National Institute of Standards and Technologies
OGC	Open Geospatial Consortium
OGF	Open Grid Forum
ONS	UK Office of National Statistics
OSA	Obstructive Sleep Apnoea
OSM	Open Street Maps
PII	Personally Identifiable Information
POSIX	Portable Operating System Interface
QA	Quality Assurance
RDBMS	Relational Database Management System
RSSC	Respiratory Support and Sleep Centre
SNMP	Simple Network Management Protocol
T.C.O	Total Cost of Ownership
URL	Universal Resource Locator
VA	Visual Analytics
VM	Virtual Machine

1 Introduction

Big Data Analytics has come a long way since first being introduced as a concept in 1997[1]. Countless tools and systems have been developed to tackle the data deluge. Many domains, such as medicine, transport and security, generate gigabytes of data on a daily basis as part of their business processes. These domains became early adopters of the technology, as the benefit of treating their data on a longitudinal basis, as opposed to case-by-case, was clear. Other domains, such as manufacturing, IT/online-retail and urban/governmental services quickly realised that their business processes were generating significant secondary (at times meta) data that could be incorporated into the organisational value chain. Thus, began the proliferation of Big Data technologies.

However, as the push for Big Data ran mostly in parallel to the advent of public cloud computing, organisations invested money into infrastructure and assets. This led to considerable vendor lock-in and limited the agility and adaptability of the organisations to adopt new technologies within Big Data sphere (e.g. machine learning). Gartner describes this as the *trough of disillusionment* that almost all technologies face after going through a *peak of inflated expectations*[2]. These problems are further compounded by the fact that Big Data software toolsets are designed around the source data, leading to a lack of interoperability, even within an organisation. This goes against the very ethos of Big Data, where being able to combine different sizes and types of datasets is paramount.

With cloud technologies maturing, and the effort of the FOSS community and several vendors, Big Data technologies are now adopting standards to ensure a modicum of interoperability across hardware, software and applications. Track&Know, with its commercial and exploitation roadmap, aims to create a reusable and replicable toolset for Big Data mobility analytics.

1.1 Mapping TRACK&KNOW Outputs

Table 1-1 maps TRACK&KNOW's Grant Agreement commitments, both within the formal Deliverable and Task description, against the project's respective outputs and work performed.

Table 1-1: Adherence to TRACK&KNOW's GA Deliverable & Tasks Descriptions

TRACK&KNOW GA Component Title	TRACK&KNOW GA Component Outline	Respective Document Chapter(s)	Justification
DELIVERABLE			
D1.3 Interoperability Requirements	*A detailed specification for interoperability, based on the NIST framework will be documented in this deliverable.*	All	This deliverable provides a detailed description of the NIST interoperability standard and how the various components of the project map to this framework.
TASKS			
	In Task 1.3 a detailed specification of a common	Chapter 4, 5, 6, 7	These chapters outline the compliance of each pilot within Track&Know to the

[1] https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/
[2] https://www.gartner.com/technology/research/hype-cycles/

Task 1.3 Interoperability Requirements	*repository schema for metadata publishing/ information sharing will be produced.*	Specifically Section 4.3	NIST framework. The final chapter describes how the technology platform maps to the same standard.
	The task will use the NIST Big Data Interoperability Framework as a reference framework.	Chapter 2, 3	These chapters outline the challenges around interoperability in Big data and describe the salient features of the NIST framework

1.2 Deliverable Overview and Report Structure

This report outlines interoperability strategies as proposed by various industrial bodies and the compliance of the project with the NIST Interoperability Framework. Through desktop research and feedback from commercial partners and practitioners the challenges involved with interoperability of bespoke analytics were identified. Using the NIST framework as key guidance and supported by components and sections of the EC-ISA^2 Report, the roadmap to truly interoperable components being delivered by Track&Know has been recorded in this report.

The next chapter (2) outlines the challenges around Big Data Interoperability and the evolution of attempts at standardisation. A brief introduction to the NIST framework is provided along with descriptions of other frameworks and standards that have evolved recently.

Chapter 3 outlines in detail the salient features of the NIST Interoperability framework and what is required to make a compliant Big Data solution.

Chapter 4 outlines the technology platform and software tools being developed. While these components will deliver the three pilots, they are being designed and developed to be reusable tooling for processing mobility data. Therefore, mapping the development strategy to the NIST framework is an important step to ensure the developed tools and systems are truly interoperable.

Chapters 5 through 7 provide a brief overview and map the compliance of the three Track&Know pilots to the NIST framework. Drawing on the specifications laid out in Deliverable 6.1, the Insurance, Medical service, and Fleet Management pilots are re-presented based on the NIST standard to ensure that methods and processes implemented are replicable and interoperable. The actual specifications of the pilots remain unchanged. It is only the problem descriptions that have been restructured to ensure a clear mapping. Deliverable 1.2 will be the next stage at which, after stakeholder involvement, the pilots may be modified or refined. The tables within Chapters 5 through 7, will be re-provided once again in D1.2 should any modifications be made.

2 Big Data Interoperability and Frameworks

At a macro level, by its very nature, Big Data problems require bespoke solutions. However, unlike holistic software solutions from the 90's and early 00's, modern software and Big Data solutions can be and should be designed and developed as modularised components. A modular development process theoretically allows for the reuse of certain modules to perform repeatable tasks while the entire processing chain is sandwiched by bespoke data handlers and pre and post-processing modules[3].

In their review paper[4] on the challenges of interoperability in Big Data, Kadadi et al., outline the seven reasons that restrict interoperability within Big Data. The infrastructural and personnel implications of introducing workflows into a business, while continuing normal operations and systems is challenging and makes interoperability difficult when **accommodating the scope of data**. Additionally, while cloud computing may solve some of the **inadequate resources** and **scalability issues**, many solutions are designed with a minimum or optimum specification in mind that may not be applicable. Heterogeneous sources of data (even just temporally heterogeneous) can lead to **data inconsistencies**, making the data error prone. A study of medical datasets revealed error rates within data of up to 26.9%[5]. Automating the correction of this data is an added complex challenge to the already complex process of integrating heterogeneous data (e.g. merging driving data (a geographically specific data point per second) with weather (temporally banded data across a geographic area) data). The **optimised queries** required to acquire the right data and the **ELT process** needed to shape it for analysis are so specific that solutions are too intrenched to be reusable – even within an organisation. Finally, the **implementation of a support system** for users of the system that can sustain the attrition of developers and power-users is a challenge that has slowed down the adoption of Big Data technologies and the reapplication of developed tools.

Track&Know aims to minimise the effect of these processes by following internationally accepted and semi-standardised processes. Interoperability within the Track&Know world extends beyond the construction of the Big Data processing software and platform. Importantly, the three pilots are also defined using Big Data interoperability standards so that future adopters can clearly identify the different modules and be able to adapt their implementation. As per the grant agreement (GA), the NIST Big Data Interoperability Framework is being utilised to define the Track&Know platform and pilots. The next section (2.1) gives an overview of the framework in the wider ecosystem of NIST, while Chapter 3 details the specifics of the Big Data interoperability framework.

During the preparation of this report, the European Commission (EC) commissioned study by PwC on interoperability was released. Published on the 31st of May 2018 for the EU-ISA[2]: Interoperability solutions for public administrations, businesses and citizens unit, this report sets out the challenges faced for Big Data interoperability in a European context and makes recommendations for improvements in technical and data standards to member states. One of the report's key focus has been on meta-data for data publishing and discovery. This facet of interoperability is not included in the NIST framework and is therefore being relied upon by Track&Know to deliver a common repository schema. A brief description of this report is included within Section 2.2.

2.1 NIST Big Data Interoperability Framework

The United States National Institute for Standards and Technologies (NIST) developed a Big Data Interoperability Framework (NBDIF) as part of its activities within the NIST Big Data Public Working Group (NBD-PWG). The first

[3] Bonner, Stephen, Ibad Kureshi, John Brennan, and Georgios Theodoropoulos. "Exploring the Evolution of Big Data Technologies." In Software Architecture for Big Data and the Cloud, pp. 253-283. 2017

[4] Kadadi, Anirudh, Rajeev Agrawal, Christopher Nyamful, and Rahman Atiq. "Challenges of data integration and interoperability in Big Data." In *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 38-40. IEEE, 2014.

[5] Goldberg, Saveli I., Andrzej Niemierko, and Alexander Turchin. "Analysis of data errors in clinical research databases." In *AMIA annual symposium proceedings*, vol. 2008, p. 242. American Medical Informatics Association, 2008.

version of this framework[6], was published in September 2015. A reviewed second edition has been published on the 28th of June 2018. Comprised of nine volumes, this interoperability framework aims to help the US derive consensus on Big Data technologies and improve interoperability. The nine volumes of the framework[7] cover the following:

- Volume 1[8]: develops a definition of Big Data and related terms necessary to lay the groundwork for discussions surrounding Big Data.
- Volume 2[9]: Big Data taxonomies developed by the NBD-PWG.
- Volume 3[10]: use cases gathered by the NBD-PWG Use Cases and Requirements Subgroup and the requirements generated from those use cases.
- Volume 4[11]: exploration of security and privacy topics with respect to Big Data.
- Volume 5[12]: Architectures White Paper Survey *(not revised in 2018)*
- Volume 6[13]: summarises the work performed by the NBD-PWG to characterise Big Data from an architecture perspective and present the NIST Big Data Reference Architecture (NBDRA) conceptual model.
- Volume 7[14]: summarises the work presented in the other six volumes, an investigation of standards related to Big Data, and an inspection of gaps in those standards.
- Volume 8[15]: uses the work performed by the NBD-PWG to identify objects instrumental for the NIST Big Data Reference Architecture (NBDRA). *(new to 2018)*
- Volume 9[16]: implementation and modernization of Big Data systems. *(new to 2018)*

Key to Track&Know is primarily Volume 6, which lays out the Reference Architecture and establishes the principles that need to be followed for interoperability. The first edition of this volume presents the key architectural components that are technology-, infrastructure-, and vendor-agnostic. The second edition (published during the preparation of this report) defines general interfaces between the above components. The third edition (expected 2021) will include validation based on implementation of use cases from Volume 3.

Fundamentally, the NBDRA establishes a common schema against which applications and frameworks (or technological components) of Big Data need to be defined. The primary actors identified by the NBDRA are the System Orchestrator, the Data Provider, the Big Data Application Provider, the Big Data Framework Provider, and the Data Consumer. Each actor either performs some processes or includes technological components. Relating any platform and application (in this case pilot), to this ontology will enable future adopters to identify where bespoke modules are required within the processing workflow. This minimises the challenges to adoption and increases future use.

The all-agnostic approach of the NIST framework and the reference architecture make this standard globally acceptable. Further by deriving its best practices from active real-world case studies, with on-going validation due in 2021, the NIST framework provides the most comprehensive industry validated approach to interoperability.

[6] Chang, Wo L., Arnab Roy, Nancy Grady, Russell Reinsch, Mark Underwood, Geoffrey Fox, David Boyd, and Gregor von Laszewski. *NIST Big Data interoperability framework*. No. Special Publication (NIST SP)-1500-1-6. 2015.
[7] https://www.nist.gov/publications/nist-big-data-interoperability-framework
[8] https://doi.org/10.6028/NIST.SP.1500-1r1
[9] https://doi.org/10.6028/NIST.SP.1500-2r1
[10] https://doi.org/10.6028/NIST.SP.1500- 3r1
[11] https://doi.org/10.6028/NIST.SP.1500-4r1
[12] http://dx.doi.org/10.6028/NIST.SP.1500-5
[13] https://doi.org/10.6028/NIST.SP.1500-6r1
[14] https://doi.org/10.6028/NIST.SP.1500-7r1
[15] https://doi.org/10.6028/NIST.SP.1500-9
[16] https://doi.org/10.6028/NIST.SP.1500-10

2.2 Competing Frameworks

The ISA[2] report for challenges and recommended practices towards Big Data Interoperability[17] approaches the problem at a lower level of abstraction compared to the NIST effort. The report is based around European Union development roadmaps such as the Digital Single Market, EU Legislation such as the General Data Protection Regulation (GDPR), and EU technologies and systems, such as EUR-Lex. However, the authors make the case that these best practice recommendations can scale beyond and form the bases of interoperability standards.

The challenges to interoperability identified by the report are: Poor data quality; Data protection considerations; Different data licences apply to the data sources; Decoupling between data producer and data user/scientist; Complex and time- consuming data integration process (Schema-level conflicts between data sources, Data-level conflicts between data sources); Increased demand for near real-time analytics requires rapid data integration; and Lack of interfacing mechanisms between systems.

Keeping these concerns in mind, the report makes recommendations of good practices for Legal, Organisational, Semantic and Technical Interoperability. The extracts below provide the best practices and recommendations at the highest level for these four facets of interoperability.

Best Practices for Legal Interoperability: when reusing data, especially in a cross-border context, it is paramount to check which legislation applies to the domain in question. For European Union Law, consult EUR-Lex[18] or use the machine-readable versions offered through CELLAR[19].

Best Practices for Organisational Interoperability: to break down data silos, organisations can invest in open (available for anyone on the web) or closed (availability within the organisation) data marketplaces or portals, where datasets can be published and discovered by potential users. Such a marketplace or 'data portal' can:

- support interoperability through a shared metadata model, like the Data Catalogue Vocabulary (DCAT)[20];
- create a single point of discoverability for data scientists
- improve data quality consistency through mutually agreed service-level agreements.

Best Practices for Semantic Interoperability: A good practice to avoid semantic gaps lies in defining how the information should be understood. This can be done on two levels. The simplest solution for this is to build shared metadata repositories that describe the content and intent of data stored in the various information systems. Another solution is to build an ontology to support interoperability. This is more challenging but allows for smooth interoperability once in place.

Best Practices for Technical Interoperability: assess providers' technology implementations for potential areas of vendor lock-in. For example, Big Data solutions are available from different vendors, including Cloudera, Horton Works, Amazon, Microsoft, IBM, etc. While are all based on the same basic open-source technology (Hadoop), they vary in technologies and versions for data integration, processing, etc.

While the recommendations for interoperability in this report are limited and Euro-centric, the sections on Organisational and Semantic interoperability strategies fill a missing voice in the discussion on interoperability. While the NIST Big Data framework is agnostic on many level and approaches the issues of interoperability at both a low and a high level, in the next chapter it becomes clear that these fundamental issues – before data can even be considered for Big Data analytics, are completely ignored. This may be reflective of the maturity level of this field within the US market (and specifically the government and military, who are the primary audience).

[17] Jens Scheerlinck, Frederik Van Eeghem, Nikolaos Loutas. "D05.02 Big Data Interoperability Analysis." PwC EU Service Study for EU-ISA[2]: Interoperability solutions for public administrations, businesses and citizens. V.1 2018
[18] http://eur-lex.europa.eu/
[19] https://data.europa.eu/euodp/data/dataset/sparql-cellar-of-the-publications-office
[20] https://www.w3.org/TR/vocab-dcat/

3 NIST Reference Architecture

The United States' National Institute of Standards and Technology's (NIST) Big Data Public Working Group (NBD-PWG) have set out a vendor-neutral, technology- and infrastructure-agnostic guidance for interoperable Big Data Infrastructures and Applications. Acknowledging that Big Data (and interoperability) is more than just a collection of technologies, the NIST Big Data Reference Architecture (NBDRA), focuses on both the information and the information technology.

The NBDRA does not provide a recommended or standardised practice for infrastructure provisioning, services or integration. It does however set out principles and ontologies to aide in ensuring interoperability.

The next section identifies the ontologies required to define a Big Data use case aiding in long term interoperability. Section 3.2 outlines the requirements of the functional components, Section 3.3 outlines the requirements for interconnections between components and section presents the NIST stance on industrial standard for the various components.

3.1 Mapping Use Cases and Requirements

The NBDRA, through a survey of current reference architectures, active use cases, and available frameworks was able to synthesise seven stages or characterisations of processes within any Big Data activity. These seven characterisations are then mapped to eight components and fabrics of the Reference Architecture (Table 3-1). This creates an ontology for future use cases to map against.

Table 3-1: Use Case Characterisation mapped to NBDRA

Use Case Characterisation Categories		Reference Architecture Components and Fabrics
Data Sources	→	Data Provider
Data Transformation	→	Big Data Application Provider
Capabilities	→	Big Data Framework Provider
Data Consumer	→	Data Consumer
Security and Privacy	→	Security and Privacy Fabric
Life Cycle Management	→	System Orchestrator; Management Fabric
Other Requirements	→	To all components and fabrics

The reference architecture combines not just the physical infrastructure and software layers (like a typical architecture) but also includes the human and application element that is typically external to the system. Table 3-2 maps out the definition of each component or fabric to the involved actors and the roles/activities at each stage. These are further divided as functional components (i.e. must perform specific tasks within a Big Data Analytics life cycle) and all encompassing (performing a global task in the background). It should be noted that the general stage of *life cycle management* can be split out into two components:

- System Orchestrator (functional component)
- Management Fabric (encompassing component)

Table 3-2: NBDRA Component Definitions and Actors

	Component/Fabric	Actors	Definition
Functional Components	System Orchestrator	Leadership, Data- Information- Software- Security- Privacy- Network- Architects	The System Orchestrator defines the systems specification and high-level (inc. governance and business requirements). This component covers both the actors and visual interface used to build the system.
	Data Provider	Enterprises, Public Agencies, Researchers and Scientists, Network Operators, End Users	Introduces data into the system. Responsible for Collecting, Persisting, and providing Transformation functions. Manages PII. Ensures data availability through programmable interfaces.
	Big Data Application Provider	Application Specialists, Platform Specialists	Handles the Collection, Preparation, Analytics, Visualisation and Access.
	Big Data Framework Provider	In-house systems, Data Centres, Cloud Providers	Operates and develops, Infrastructure frameworks, Data Platforms and Processing frameworks.
	Data Consumer	End Users, Researchers, Applications, Systems	Perform search and retrieve, download, analyse locally, reporting and visualisation processes.
Encompassing	Security and Privacy Fabric	CIO, CSIRT, Specialists	Processes and mechanisms at each stage of the infrastructure and Big Data lifecycle, ensuring security and privacy.
	Management Fabric	In-house Staff, Data Centre Management, Cloud Providers	Covers both the management of the infrastructure using manual, semi- and automated tools (e.g. SNMP), and the Big Data lifecycle.
	Other Requirements	End Users, System Designers	Incorporating emerging technologies and improving usability (e.g. IoT support, mobile interfaces)

3.2 Conceptual Model and Infrastructural Layering

The NIST Big Data Reference Architecture (depicted in Figure 3.1) layers the Big Data Architecture along two axes: the Information Value Chain and the IT Value Chain. The *blue* arrows show data flows between components and owners, while the *red* arrows depict the software tools and algorithms that perform conversions to make data available and correctly formatted for the receiving component. The conversion tools run in-situ to the source of the data before transferring. The green arrows depict the initiating source within the Big Data lifecycle (i.e. Data Consumer calls on the Big Data Application Provider, which calls on the Data Provider).

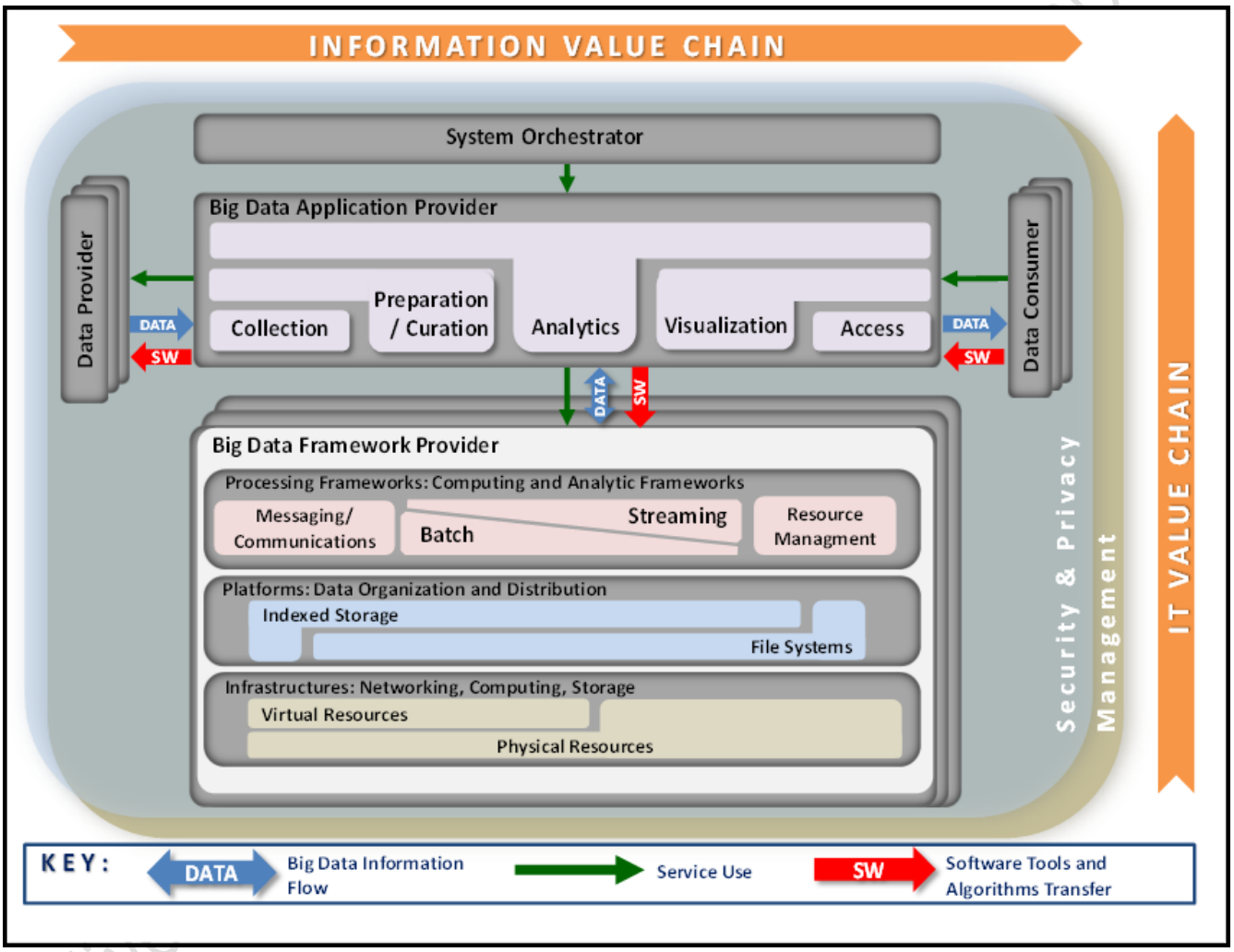

Figure 3.1: NIST Big Data Reference Architecture[21]

Along the Information Value Chain, value is derived from raw data through Big Data operations. The IT Value Chain is typical to most system architectural diagrams outlining the hierarchy of technology, network at the bottom, infrastructure and platforms in the middle, and applications on top.

The Pilots (Work Package 6) within Track&Know span across the Informational Value Chain, while the development of tools and the platform (Work Packages 2-5) go up the IT Value Chain. The NBDRA further breaks

[21] Chang, Wo L., Arnab Roy, Nancy Grady, Russell Reinsch, Mark Underwood, Geoffrey Fox, David Boyd, and Gregor von Laszewski. *NIST Big Data interoperability framework*. No.1 Special Publication (NIST SP)-1500-1-6. pp 11. 2015.

down the Big Data Application Provider and Big Data Framework Providers. These are discussed in the next two sections.

3.2.1 Big Data Application Provider

The Big Data Application Provider (BDAP) performs five types of operations along the data life cycle within the Architecture. Dictated by the design of the System Orchestrator, the BDAP can include several components typically seen in traditional systems. However, it is the *volume, variety and velocity* of the data that makes these components or stages of the pipeline unique.

The different operations that these components carry out are:

Collection: Within the BDAP the collection activity is the automated aggregation of data on to a centralised platform for processing. Typically, these are *pull* activities where the platform is designed to periodically connect to an external source over the internet and retrieve the data e.g. social media data. *Push* activities are also possible where end points connect to the BDAP to transfer data. These are typically seen in either low bandwidth or batch import environments. These interfaces are standardisable and data collection assets can be reusable.

Preparation: The Extract Load and Transform (ELT) operations required to shape the collected data for either processing or efficient storage within the infrastructure. Usually, this is a custom process that tends not to be migratable as it is the missing link in merging external and internal assets.

Analytics: The specific data processing algorithms for processing the data to produce new insights that will address the technical goal of the Big data lifecycle form the analytics components. With properly documented interfaces and good error handling Analytics components are interchangeable and reusable. For improved interoperability, the specifics of the Analytics components need to be disclosed. This will increase System Orchestrator confidence in the processes and help in tool adoption decision making (e.g. training set for a deep learning module).

Visualisation: This is the process to transform the data (raw or post ELT) and the results of the analytics into a format that is able to optimally communicate meaning and knowledge. Within the NBDRA even text reports are included as visualisations. Highly modular, visualisation components are interchangeable and can form frameworks in their own right.

Access: Differing from the security layer, the access component of the BDAP is the interface layer that allows the data consumer to get access to the analytics and visualisation. Additionally, the data consumer is able to use the access component to start new life cycles along the *Information Value Chain*.

The NBDRA recognises that activities within this layer would be application specific and therefore are not candidates for standardisation, however, efforts should be made to ensure that the metadata and policies defined and exchanged between subcomponents are standardised specific to the vertical industry.

3.2.2 Big Data Framework Provider

The Big Data Framework Provider comprises the following three subcomponents or services.

Infrastructure Frameworks: At the lowest (closest to the electronics) are the infrastructure frameworks. These frameworks including the Networking (cabling, switches and routers etc.), Computing (processors, memory, accelerators), Storage (in-situ, centralised devices), and Environment (power, cooling etc.) elements. Many of these elements may not be relevant to all adopters of Big Data solutions. Due to virtualisation, the networking and computing elements can be aggregated, and the appliances are agnostic of the system below (implied interoperability). Cloud computing and containerisations further enhance the interoperability.

Data Platform Frameworks: Sitting above the Infrastructure framework is the Data platform. The Data Platform Framework goes beyond the storage elements found in Infrastructure Frameworks. A data platform is responsible for the ingestion, indexing, storage and recall of data. This is a critical component of any Big Data system as the volumes and the divide-and-conquer approaches to processing data have significant management

implications beyond file reads and writes. Different problems require the data to be arranged and made available in different structures. Where possible in-memory processing i.e. the loading of data in the systems main memory, if preferred as processing times are reduced. However, the size and divisibility of the data will determine whether these techniques are possible. There is also a data transfer cost that needs to be considered. The cost implications further change depending on the method of data storage within the file system. The NIST framework does specify that based on the survey a Portable Operating System Interface (POSIX) compliant file system is and should be the preferred method. Due to the volume and variety of data the NIST framework also recommends an indexed storage approach at the Data Platform level. Some method of either key-value stores, relational, graph or document-model based indexing needs to be employed when storing data to improve data accessibility and traceability.

Processing Frameworks: The top most layer of the Big Data Framework Provider is the Processing Framework. This layer sits between the hardware and its associated management layers and the algorithms and tools found in Big Data Application Provider layer. Depending on the type of execution and analysis, the processing frameworks fall under a combination of batch frameworks, streaming frameworks, communication frameworks, and resource management frameworks. Where there are large quantities of data at rest, batch frameworks and communications frameworks such as Map-Reduce, Bulk Synchronous Parallel (BSP) and Message Passing can be utilised to process the data longitudinally. For on-line systems where data is constantly entering the system Stream computing is used. The NBDRA recognises that current implementations of stream processing employ some form of Directed Execution and Access Graphs (DEG/DAG). Resource Management Frameworks are those that can dynamically adjust the execution environment to meet the needs of the processing workflows. In terms of a cloud computing environment the resource management framework ensures an elastic scale up of the processing systems.

3.3 Interconnections

3.3.1 High Level Principles

With Version 2 of Volume 6 and the introduction of Volume 8, of the NBDRA, the principles governing the interconnections between the high-level and low-level components have also been specified. Overall the reference architecture acknowledges that at every level there are multiple compliant solutions and products. To meet the interoperability challenge, all implementations need to ensure Technology- and Vendor- Agnostic interfaces (i.e. do not use proprietary pay-wall blocked formats and technologies), the support for a plug-in infrastructure, and the orchestration of infrastructure, services, applications and experiments.

Plug-in infrastructure refers to solutions not being specification locked, i.e. if a high-memory system is not available then the tooling needs to adjust to what is available, similarly, the availability of accelerators should factor into the execution. There is no requirement for this support to be automated but, as an example, be made possible through a code recompile. Further the servitisation of analysis components can further increase interoperability, e.g. provision of an image classifier on GPU's in the cloud and open an API for laptops or mobile devices to connect to. The architecture also encourages the creation of execution payloads that are able to co-locate with the data to perform the analysis in-situ. This is especially important if the data is too large or there are confidentiality or legality concerns (the NIST use-cases that the reference architecture is based on are primarily for the U.S. government and military thus the lean towards confidentiality concerns).

Orchestration is an important facet for reusability purposes. Being able to script not only the experiment but also the software and application execution environment, ensures that downstream adopters are available to migrate solutions and determine processes for interoperability.

3.3.2 Component-Specific Interconnections

The general principles specified in the section above apply to the seven layers within the reference architecture, but the framework makes specific recommendations or refers to the use of specific open technologies, which conform to accepted standards (Discussed in Section 3.4), for certain layers. For other layers it identifies challenge areas.

For the Data Provider, backward compatibility of their API's with historic datastores are essential. As data gathering and storage paradigms evolve it may become difficult or costly to maintain API's. The interoperability implications require either data migration or new layers to connect new access mechanisms to the historic datasets.

For the Big Data Framework Provider, layer interfaces must ensure users access to the underlying data. This layer must provide interfaces to files, interfaces to virtual data directories, interfaces to data streams, and interfaces to data filters.

3.4 Standards

The NBDRA acknowledges that there are presently many approaches and standards towards the various components and layers of any Big Data implementation. Different layers of the reference architecture have different level of maturity and therefore different standards. Among others the NBDRA (in Volume 7, 1st edition) recognises the following standards as acceptable for the development of components within the architecture:

- International Committee for Information Technology Standards (INCITS) and International Organization for Standardization (ISO)—de jure standards process
- Institute of Electrical and Electronics Engineers (IEEE)—de jure standards process
- International Electrotechnical Commission (IEC)
- Internet Engineering Task Force (IETF)
- World Wide Web Consortium (W3C)—Industry consortium
- Open Geospatial Consortium (OGC®)—Industry consortium
- Organization for the Advancement of Structured Information Standards (OASIS)—Industry consortium
- Open Grid Forum (OGF)—Industry consortium

The second edition of Volume 7 breaks down several of the technologically mature layers and provides, where available, specific standards (e.g. ISO standards). However, the report outlines that this is an evolving list, which will be further refined in the third edition.

4 Technology Platform

4.1 Outline

Figure 4.1: Track&Know High-Level Architecture presents the various components of the Track&Know Platform Architecture. This is the preliminary architecture that will be refined and formalised in D1.2. In that deliverable a re-evaluation of the various components will be performed with respect to the NIST Big Data Interoperability Framework. Further details of the initial version can be found in D6.1.

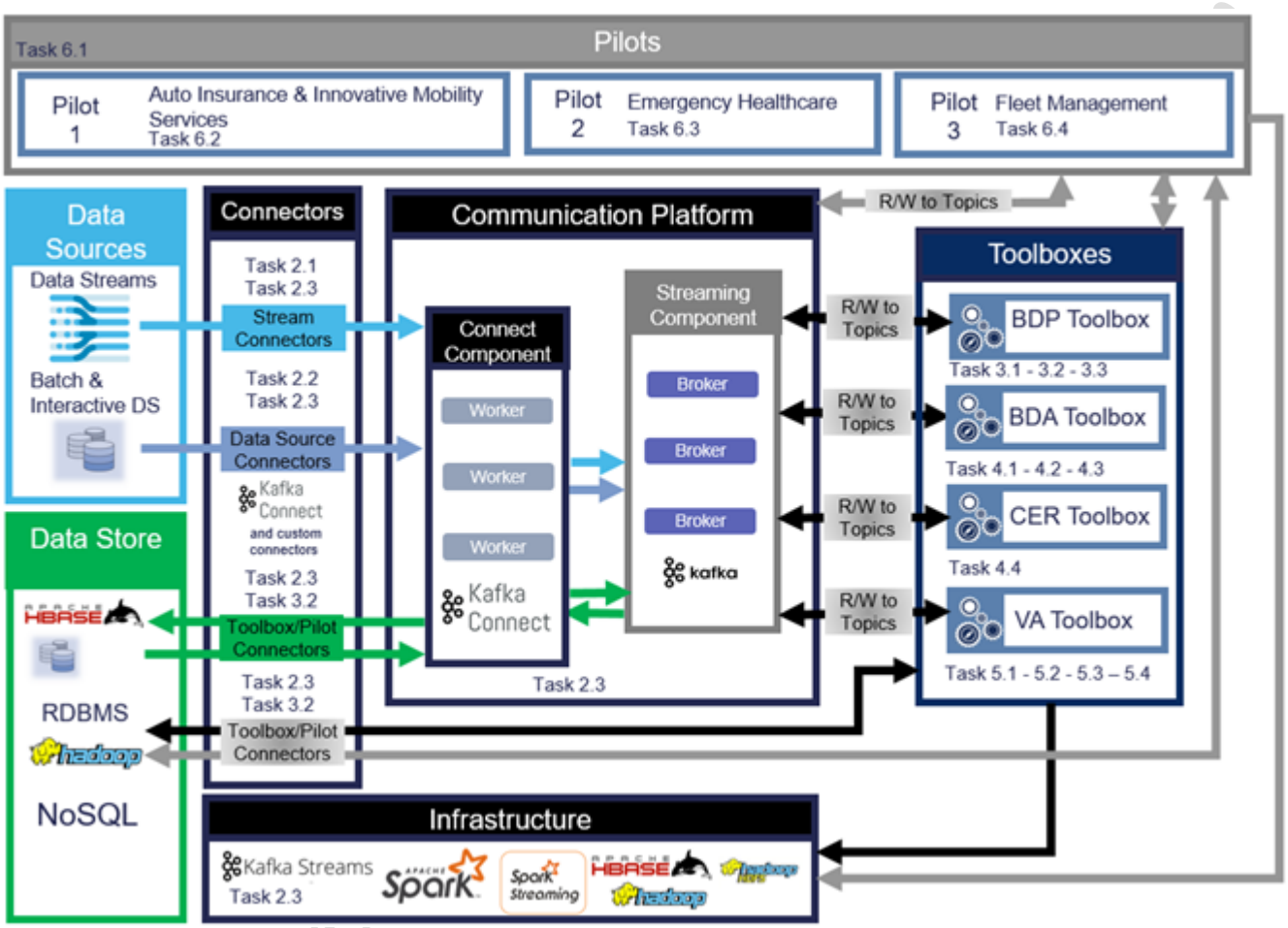

Figure 4.1: Track&Know High-Level Architecture

The Track&Know Architecture comprises of 5 component layers:

Data Sources and Data Store: The external to the platform input data sources (Data Sources) represent the structured and unstructured data streams and the batch and interactive data sources that will be made available and will be connected to the platform.

Connectors and Communication Platform: External data sources will be connected and made available by employing the *Stream Connectors* and *Data Source Connectors*. The Communication Platform represents a distributed streaming platform, which is an Apache Kafka cluster.

Toolboxes: The developed toolboxes deliver the analytics to the end user and consume incoming data via the Communication Platform. Toolbox code will therefore be able to connect to a Streaming Component Broker and consume data from a Topic. Similarly, Toolbox code will be able to publish data to a Topic (that can be further used by e.g. another job, Toolbox or a Pilot).

Pilots: The Pilot implementations will indirectly access and process incoming and existing data by using the Toolboxes. Furthermore, a Pilot implementation will be expected to consume derived data that the Toolboxes' processing will be producing.

Infrastructure: The underlying infrastructure will span across multiple virtual machines VMs and will provide all the necessary technologies and components that will enable the storage and analysis of the data involved, allow the usage of Toolboxes and facilitate the execution of Pilots, by providing a distributed computing environment that enables the above.

4.2 Interoperability Mapping

Architecturally, the various layers and components described above reflect a logical break-down of the consortium's implementation plan. While there is no 1-to-1 relationship between the layers, Table 4-1 demonstrates how each component from Figure 4.1 within the layers of the Track&Know platform (represented column-wise) map to the layers of the NBDRA (row-wise). For the purposes of this report, it should be noted that there is an implicit layer of security and privacy management between the various sources and sinks in the system. This will be further elaborated upon during the final system design in D1.2. It should also be noted here that currently privacy to a large extent is a manual process. Due to the nature of the datasets, the multi-national actors within the consortium, and the passing of GDPR regulation during the preparation of this report, the consortium is taking a cautious approach to privacy and the data providers are anonymising data at source.

Table 4-1: Component to NIST Architecture Layer Mapping

NBDRA\T&K	Pilots	Data Sources	Data Store	Communication and Connectors	Toolboxes	Infrastructure
System Orchestrator						System and Visualisation Front End.
Data Provider		Pilot 1, Pilot 2, Pilot 3, Weather, Traffic, Demographic Data,				
Big Data Application Provider					BDA, BDP, CER, VA Toolboxes	System and Visualisation Front End.
Big Data Framework Provider			RDBMS, Hadoop, HBASE, NoSQL	Stream, Data Source, Toolbox Connectors, Kafka Connect		Kafka Stream, Hadoop, Spark, Spark Streams
Data Consumer	Pilot 1, Pilot 2, Pilot 3					
Management Fabric						Platform Management Interface

4.3 Common Repository Schema

As previously mentioned in Section 2.2, the NIST framework and the NBDRA does not give guidance on data level interoperability. This report believes that this is in part due to the intended target audience of NIST – the U.S. government. The U.S. has led in the space of Big Data within the public sector for a long time and has invested billions in data aggregation, culminating in the creation of the U.S. Data Federation to support government-wide standardisation and federation[22].

The EU has now put considerable effort in data integration and standardisation and the framework put forward by ISA[2] in their report on recommended practices towards Big Data Interoperability[23] fills the missing gap faced by Track&Know in guaranteeing interoperability compliance. Chapter 3.3 in the ISA[2] report on *Closing the Semantic Gap between Datasets* outlines several good practices and sources for approved common schema for data-sharing.

The report maps out several semantic libraries for system developers to consult when publishing data. In the absence of vocabulary, for a particular industry vertical, the report gives best practices for defining a proprietary vocabulary. The report however does stress that community supported vocabularies that can be reused for a specific domain or interest should be given emphasis and preference. If the vocabulary is weak or incomplete projects and industry supporting its development (and not going proprietary) will help long term interoperability. The report identifies the following resources for existing and mature semantic vocabulary:

- ISA[2] Core Vocabularies[24]
- Linked Open Vocabularies portal[25]
- List of W3C endorsed vocabularies[26]
- Eurovoc, the EU's multilingual thesaurus[27]
- European Publication Office's Named Authority Lists[28]
- Data Cube Vocabulary[29]
- W3C's DCAT

It should be noted here that the ISA[2] Core Vocabularies semantic library has been accredited as W3C compliant as it is based on standards such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL). Invariably, this also make the ISA[2] Interoperability framework and its approach to dataset semantic schemas NIST compliant. The ISA[2] Core Vocabularies also provides detailed guidance to enhance and contribute to its semantic vocabulary[30].

As data is integrated into the Track&Know platform (or case permitting) where the dataset is opened by the Data Provider the following sections of the ISA[2] Core Vocabulary will be employed to store the data:

- **The Core Location Vocabulary:** a simplified, reusable and extensible data model that captures the fundamental characteristics of a location, represented as an address, a geographic name, or geometry.
- **The Core Person Vocabulary:** a simplified, reusable and extensible data model that captures the fundamental characteristics of a person.

[22] https://federation.data.gov/about/
[23] Jens Scheerlinck, Frederik Van Eeghem, Nikolaos Loutas. "D05.02 Big Data Interoperability Analysis." PwC EU Service Study for EU-ISA[2]: Interoperability solutions for public administrations, businesses and citizens. V.1 2018
[24] https://joinup.ec.europa.eu/page/core-vocabularies
[25] http://lov.okfn.org/dataset/lov/
[26] https://www.w3.org/standards/techs/rdfvocab
[27] http://eurovoc.europa.eu/
[28] http://publications.europa.eu/mdr/authority/index.html
[29] https://www.w3.org/TR/vocab-data-cube/
[30] https://joinup.ec.europa.eu/site/core_vocabularies/Core_Vocabularies_user_handbook/ISA Hanbook for using Core Vocabularies.pdf

- **The Core Business Vocabulary:** a simplified, reusable and extensible data model that captures the fundamental characteristics of a legal entity.

It is envisaged that Track&Know will recommended contributions to these standard libraries.

4.4 Ensuring Compliance

Track&Know aims to use Kafka Connectors and Topics as its communication layer between the various components of the platform, and base its processing on Spark, Hadoop and Kafka platforms. These are all projects of the Apache Foundation. The Apache Foundation itself, is W3C, IETF and OGF compliant in its development. Therefore, by planning Track&Know's with the Kafka and Spark ecosystems compatibility, the project will ensure a NIST compliant implementation.

In order to address poor data quality outlined as a challenge in the relative ISA report on Big Data challenges[31] and recommended practices the proposed architecture can accommodate within its pipelines the necessary data cleansing and validation approaches which when introduced in established data flows can further guarantee the data quality. As far as Data protection considerations are concerned, the technologies selected offer useful tools towards ensuring the protection of data while in transit and at rest. More specifically, the encryption of data in-flight can be achieved using TLS which allows data to be encrypted between your producers, consumers and the Kafka Cluster. Authentication is in turn achieved using SSL or SASL which allows producers consumers to authenticate verifying thus their identity. If a more fine-grained approach to Authorization is necessary, it can be achieved by using access control lists (ACLs). Authenticated clients can be run against access control lists to determine whether a client can proceed to use specific topics (read and/or write).

It is worth mentioning that the Decoupling between the data producers and data users is guaranteed by the technologies involved and facilitated by the existence of a. Open source libraries for producer and consumer APIs for almost all the important programming languages, b. The utilisation of a High-Performance Broker approach (Kafka) which serves as the main de-coupling component between producers and consumers and c. The straightforward, production and consumption of data in a publish-subscribe fashion and over encrypted communications when necessary.

In order to avoid complex and time- consuming data integration processes which include Schema-level conflicts between data sources and Data-level conflicts between data sources, the proposed approach is capable of using JSON Schemas for the data and furthermore can provide its Schema Registry functionality as a serving layer for metadata. The Registry provides a RESTful interface for storing and retrieving Avro schemas while maintaining a versioned schema history and allowing evolution of schemas according to the configured compatibility settings and expanded Avro support. The Schema Registry provides serializers that Kafka clients can use to handle schema storage and retrieval for Kafka messages sent in the Avro format.

The rapid data integration requirement originating from the increasing demand for near real-time analytics can be adequately addressed by the Kafka technologies which can accommodate end-to-end real-time data pipelines that can scale out according to needs and increasing loads. The technologies allow for multiple producers and consumers of data to operate in parallel, allowing designs that effectively partition the data and its processing as necessary ultimately achieving the required system responsiveness.

In its entirety, the Kafka Cluster can be considered also as an efficient interfacing mechanism between systems, capable of integrating various components and advancing the overall interoperability.

[31] Jens Scheerlinck, Frederik Van Eeghem, Nikolaos Loutas. "D05.02 Big Data Interoperability Analysis." PwC EU Service Study for EU-ISA[2]: Interoperability solutions for public administrations, businesses and citizens. V.1 2018

5 Pilot 1: Insurance

5.1 Outline

The auto-insurance use case led by Sistematica (P-15), will utilise data from vehicle black boxes (provided by Octo Telematics) to gain insights for the following three insurance sector business questions[32].

IUC-1 - Insurance Business Case: Using historic telematics, environmental, demographic and geographic information to answer the questions "What is the probability that driver X will have a car crash within the N kilometres". It is of the industries' interest to gain in-depth and accurate crash probability estimation. In order to tailor a driver's potential driving risk, it is mandatory to know as much as possible about the riskiness of the driving condition and the dangerous behaviours that the driver is used to display while driving under those conditions.

Therefore, the Track&Know goal is to extract relevant semantics from data collected through telematics devices in order to infer driver's habits, driving patterns, risky behaviour on the road. A parameter that describes the likelihood of a car crash occurring given certain condition and how this parameter could be used to guide a driver's behaviour in order to decrease driving risk conditions and consequently decrease insurance fees, is the main KPI for this use case. Globally the Track&Know platform needs to support semantic extraction, data enrichment/fusion, pattern recognition, clustering and stratification and model performance operations to deliver the use case.

IUC-2 - Electric Car Mobility Case: The adoption of electric vehicles (EV) will help lower total cost of ownership (T.C.O) of vehicles for consumers. However, before a switch can be made the following questions need to be answered: "Will the driver experience any considerable benefits by switching from his/her mobility paradigm to an electric car mobility?" and "Will the available charging points match customers' needs in terms of charging time and geographical locations?".

This use case aims to derive two parameters. The first being a parameter that describes cost-benefit of a switching to an electric car mobility with respect to traditional mechanical engines, and the second parameter that describes how global charging times and charging points match the driver's habits. Globally the Track&Know platform needs to support clustering and stratification, and cost estimation operations to deliver this use case.

IUC-3 - Car Pooling Business Case: To reduce urban congestions by increasing the number of users utilising a Car-pooling service, it is important to answer two questions: "Will both the driver and the passenger experience any considerable benefit by sharing a ride?" and "Will a carpooling service meet passengers' demand in terms of time/zone availability and time-matching?" These business questions result in statistical parameters that are derived from:

a) An estimation of circulating park decreasing due to sharable routes in the three areas considered with a consequent decreasing of global insurance risk.
b) A parameter that describes cost-benefit of switching to a sharing mobility paradigm instead of owning a car
c) A parameter that describes the likelihood of finding a proper sharable route that matches time and geographical zone and shall easily substitute the usual routes travelled.

Globally the Track&Know Platform will need to support analysis that performs Clustering and Stratification to deliver the above investigations.

5.2 Interoperability Mapping

In Table 5-1, the three use cases and their various components are mapped to NBDRA.

[32] Further details can be found in Chapter 2 of D6.1 Experiments Planning and Setup

Table 5-1: Insurance Pilot: Business Question NBDRA Mapping

	Data Provider	Big Data Application Provider	Big Data Framework Provider	Data Consumer
IUC-1	OCTO Telematics API, Weather API,	BDP, BDA, CER, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics	SIS/CNR
IUC-2	OCTO Telematics API, EV Network API (e.g. Zap-Map)	BDP, BDA, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics	SIS/CNR
IUC-3	OCTO Telematics API	BDP, BDA, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics	SIS/CNR

6 Pilot 2: Healthcare Service

6.1 Outline

Royal Papworth Hospital (PAP) has provided sleep services since 1991. The Respiratory Support and Sleep Centre (RSSC) is the largest of its kind in the UK and was the first, and remains one of only two, to be accredited by the European and British Sleep Societies as a mark of the quality of its service. The service is under stress due to demand, yet many suffers of Obstructive Sleep Apnoea (OSA) remain undiagnosed or untreated, potentially in part due to the distances involved in traveling from rural settings to the Hospital. PAP aims to increase the outreach and effectiveness of their service while ensuring it continues to meet its target of 18 weeks from referral to first attempt at treatment. This target will remain achievable by ensuring the diagnostic devices reach the patients in the most cost-effective manner and matches the propensity maps for this condition. There are 6 objectives this pilot would like to address[33].

HCUC-1 – Improvement of response times (increasing new patients): Demonstrate a higher throughput of new patients in the service keeping to the 18-week referral.

HCUC-2 – Improvement of response times (increasing follow-up patients): Demonstrate a higher throughput of follow-up patients in the service without purchasing more diagnostic equipment.

HCUC-3 – Reduction of unnecessary travel (reduce patient travel distances): Demonstrate and visualize a summative reduction of travel.

HCUC-4 – Reduction of unnecessary travel (reduce courier costs): Demonstrate and visualize a reduction of courier journeys, which are charged by distance, required to service the Oximetry for the number of patients over a period of time.

HCUC-5 – Reduction of unnecessary travel (reduce CO2 emissions): Simulation of CO2 reduction after demonstration of the changes in travel distance at baseline and after HCUC-4.

HCUC-6 – Cost efficiency gains: Reduce total costs for the OSA service per patient undergoing oximetry by performing total cost calculations per simulation scenario in the above use case scenarios.

Globally the Track&Know platform needs to support prediction, cost estimation, data integration and fusion, and simulation operations to deliver the above use cases.

6.2 Interoperability Mapping

In Table 6-1, the six use cases within this pilot and their various components are mapped to NBDRA.

[33] Further details can be found in Chapter 3 of D6.1 Experiments Planning and Setup

Table 6-1: Healthcare Service Pilot: Business Question NBDRA Mapping

	Data Provider	Big Data Application Provider	Big Data Framework Provider	Data Consumer
HCUC-1	NHS-PAP: (Patient Scheduling, GP Information Patient Information and Propensity Maps), ONS (Demographics)	OSM Geocoder, BDP, BDA, VA	BMDI: NoSQL, HBASE, Containerised Compute	PAP/CEL
HCUC-2	NHS-PAP (Patient Scheduling, GP Information Patient Information and Propensity Maps)	OSM Geocoder, BDP, BDA, VA	BMDI: NoSQL, HBASE, Containerised Compute	PAP/CEL
HCUC-3	NHS-PAP: (Patient Scheduling, GP Information Patient Information and Propensity Maps, Timetables), ONS (Demographics), OpenStreetMaps, Store API's	OSM Geocoder, BDP, VA	BMDI: NoSQL, HBASE, Containerised Compute	PAP/CEL
HCUC-4	NHS-PAP: (Patient Scheduling, GP Information Patient Information and Propensity Maps, Timetables), ONS (Demographics), OpenStreetMaps, Store API's	OSM Geocoder, BDP, BDA, CER, VA	BMDI: NoSQL, HBASE, Containerised Compute	PAP/CEL
HCUC-5	NHS-PAP: (Patient Scheduling, GP Information Patient Information and Propensity Maps, Timetables), ONS (Demographics), OpenStreetMaps, Store API's	OSM Geocoder, BDP, CER, VA	BMDI: NoSQL, HBASE, Containerised Compute	PAP/CEL
HCUC-6	NHS-PAP: (Patient Scheduling, GP Information Patient Information and Propensity Maps, Timetables, costs), ONS (Demographics)	OSM Geocoder, BDP, CER, VA	BMDI: NoSQL, HBASE, Containerised Compute	PAP/CEL

7 Pilot 3: Fleet Management

7.1 Outline

Utilising the Track & Know developments and the Big Data Platform, this pilot aims to improve the processes and commercial services of the Fleet Operator, Vodaphone Innovo (ZEL). Platform tools developed in other Pilots will be employed in the set-up, implementation and execution of the pilot, in an attempt to promote the take-up of advanced Big Data technologies in commercial Fleet Management applications. The Track & Know platform and data-driven toolboxes will allow the acquisition of data from the Fleet Management Data Source, filter and aggregate them before analysing them in a performant way, implementing thus the Fleet Management pilot.

The overall goal of improved services will be delivered by determining Big Data solutions to the following problems[34]:

FMUC-1 - Predictive maintenance

FMUC-2 - Anomaly detection, reduction of false alarms

FMUC-3 - Correlation of Fleet Data with external Weather and Traffic services

FMUC-4 - Fleet costs reduction

FMUC-5 - Fleet downtime reduction

FMUC-6 - Fleet response time improvement

FMUC-7 - Improve driver behaviour and reduce accidents

7.2 Interoperability Mapping

In Table 7-1, the seven use cases within this pilot and their various components are mapped to NBDRA.

[34] Further details can be found in Chapter 4 of D6.1 Experiments Planning and Setup

Table 7-1: Fleet Management Pilot: Business Question NBDRA Mapping

	Data Provider	Big Data Application Provider	Big Data Framework Provider	Data Consumer
FMUC-1	ZEL (Historic & Streaming API), ZEL (Maintenance and Driver Records)	BDP, BDA, CER, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics, Kafka/Spark Streaming – Cloud Hosted	ZEL
FMUC-2	ZEL (Historic & Streaming API)	BDP, BDA, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics, Kafka/Spark Streaming – Cloud Hosted	ZEL
FMUC-3	ZEL (Historic & Streaming API), Weather API, Traffic API, ZEL (Driver Records)	BDP, BDA, CER, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics, Kafka/Spark Streaming – Cloud Hosted	ZEL
FMUC-4	ZEL (Historic & Streaming API)	BDP, BDA, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics, Kafka/Spark Streaming – Cloud Hosted	ZEL
FMUC-5	ZEL (Historic API), ZEL (Maintenance and Records)	BDP, BDA, CER, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics – Cloud Hosted	ZEL
FMUC-6	ZEL (Historic API), ZEL (Maintenance Records)	BDP, BDA, CER, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics – Cloud Hosted	ZEL
FMUC-7	ZEL (Historic API), ZEL (Driver Records)	BDP, BDA, CER, VA	BMDI: Hadoop/Spark Framework, HBASE, Kafka Topics – Cloud Hosted	ZEL

8 Conclusions

This report aimed to demonstrate the Track&Know consortium effort at ensuring the interoperability and re-use of components designed and developed within the life of this project. By paying close attention to the developing field of Interoperability standards and frameworks, future deliverables, e.g. D1.2, will continue the effort of relating all activities to known best practices. In particular, the Track&Know project aims to ensure compatibility with the United States' National Institute of Standards and Technologies' Big Data Interoperability Framework. Surveying the available frameworks and standards, the NIST Big Data Reference Architecture is the most mature and comprehensive.

However, frameworks are continuing to develop and adapt. During the preparation of this report both the EU and the U.S. have released new or enhanced interoperability frameworks, further complicating efforts towards interoperability. The consortium will continue to pay close attention to the development of the NIST framework, particularly with respect to the technology layer, and once the platform is finalised, will ensure component level compliance. The expected roadmap of the NBDRA will result in the final revision of the framework to be published after the life of the Track&Know project. However, it is expected that the fundamental descriptions and layers will not change significantly.

Annex I: Ethics Proforma

Ethics Check-List and Questionnaire

A. PERSONAL DATA

No Personal Data is processed for this deliverable.

1. Has personal data going to be processed for the completion of this deliverable?
    1. If "yes", do they refer only to individuals connected to project partners? Or to third parties as well?

Not Applicable

2. Are "special categories of personal data" going to be processed for this deliverable? (whereby these include personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, and trade union membership, as well as, genetic data, biometric data, data concerning health or data concerning a natural person's sex life or sexual orientation)

Not Applicable

3. Has the consent of the individuals concerned been acquired prior to the processing of their personal data?
    1. If "yes", based on the Project's Consent Form? On a different legal basis?

Not Applicable

4. In the event of processing of personal data, is the processing:
    1. "Fair and lawful", meaning executed in a fair manner and following consent of the individuals concerned?
    2. Performed for a specific (project-related) cause only?
    3. Executed on the basis of the principle of proportionality (meaning that only data that are necessary for the processing purposes are being processed)?
    4. Based on high-quality personal data?

Not Applicable

5. Are all lawful requirements for the processing of the data (for example, notification of the competent Data Protection Authority(s), if applicable) adhered to?

Not Applicable

6. Have individuals been made aware of their rights (particularly the rights to access, rectify and delete the data)?

Not Applicable

B. DATA SECURITY

1. Have proportionate security measures been undertaken for protection of the data, taking into account project requirements and the nature of the data?
    1. Brief description of such measures (including physical-world measures, if any)

There are confidentiality concerns as this report is a public document. Approval for the release of this from the various Pilot partners who may have confidentiality concerns.

2. Is there a data breach notification policy in place within your organisation?

Not Applicable

C. DATA TRANSFERS

No Data transfers were required for the preparation of this deliverable.

1. Are personal data transfers beyond project partners going to take place for this deliverable?
    1. If "yes", do these include transfers to third (non-EU) countries?

Not Applicable

2. Are personal data transfers to public authorities going to take place for this deliverable?
    1. Do any state authorities have direct or indirect access to personal data processed for this deliverable?

Not Applicable

3. Taking into account that the Project Coordinator is the "controller" of the processing and that all other project partners involved in this deliverable are "processors" within the same contexts, are there any other personal data processing roles attributed to any third parties for this deliverable?

Not Applicable

D. ETHICS AND RELATED ISSUES

1. Are personal data of children going to be processed for this deliverable? **No.**
2. Is profiling in any way enabled or facilitated for this deliverable? **Partial Facilitation. This report only outlines the use of automated profilers. Specific details will be finalised and sent for ethical clearance w/ D1.2 and d6.2, d6.3, d6.4.**
3. Are automated-decisions made or enabled for this deliverable? **No.**
4. Have partners for this deliverable taken into consideration system architectures of privacy by design and/or privacy by default, as appropriate? **Yes.**
5. Have partners for this deliverable taken into consideration gender equality policies? **Not Applicable**
6. Have partners for this deliverable taken into consideration confidentiality of the data requirements? **Yes.**