Decentralized Data Marketplace

FROM DIGITIZATION TO DATAFICATION
EXTRACTING MEANINGFUL STORIES
FROM HETEROGENEOUS DATASETS

Paris, June 2019

BRAINCITIES

Cross-Curated Data Framework (CCDF) for A Decentralized Peer to Peer Data Marketplaces

Rishav Anand
BRAINCITIES LAB
rishav.anand@braincities.co

Kazé A. ONGUENE
BRAINCITIES LAB
contact@braincities.com

Daniel S COVACICH
BRAINCITIES LAB
contact@braincities.com

Abstract — Even if initiatives like open-data are getting more and more adopted at the governmental level, access to most of the data produced by our increasingly digitized environment remain extremely limited due to its proprietary nature. BRAINCITIES LAB aims to solve this problem by developing a peer-to-peer marketplace providing a single point of access for open and proprietary datasets. Thereby data would be leveraged for the development and implementation of unlimited number of new digital services and the forecasting of events that today are unpredictable. This paper presents the architecture for a peer-to-peer Data Marketplace, an enabler for cross-sectorial and innovative digital services platform. The Cross-Curated Data Framework (CCDF) is defined as an open and harmonized data standard, allowing the aggregation and interoperability of heterogeneous datasets. This paper explore the conditions of deployment of such advanced CCDF and the implementation of a peer-to-peer Data Marketplace at a governmental scale. The two use-cases of local weather prediction and road quality measurements are introduced to show the applicability of the Cross-Curated Data Framework (an industry agnostic data refinery) and prototype to a variety of applications.

Index Terms—*Big Data, Big Data indicators, Big Data measurement,, information infrastructure, Datafication, Decentralized Data Exchange, human factor.*

I. INTRODUCTION

Modern economy and ecosystems like cities rely on intertwined networks of sensors and technologies gathering billions of signals to be processed by expert systems in quasi-real-time. Still, traditional players are focused on the analysis of these data extracted by sensors intended to control and safety related management tasks. Rising *Advanced Decision Making Systems (ADMS)* are pushing forward in new dimensions of customer satisfaction, people happiness as well as AI augmented decision making.

To tackle highly technical challenges like autonomous driving, the dominant trend tends to favor the increase of installed sensors and complex signal processing units. We see rooms for improvement in these millions of connected cars roaming Dubai's streets. It creates a huge potential for applications and services build upon in-aggregated data to combine with environmental data extracted by traffic monitoring systems and soon street embedded sensors. All This potential is still locked since industries were siloed, unable to create a sustainable and interoperable open data based ecosystem. Inspired by how information and knowledge are shared by academics, this paper proposes an exchange platform for preprocessed and curated data taking the form of a Decentralized Big Data Marketplace enabling the development of new digital services for any industry.

Fig. 1 shows the processing chain for Big Data aggregation. Phone owners are used as data miners to collect information about themselves and their environment using phone built-in sensors. Gathered data are transferred into the cloud leveraging phone-to-cloud communication systems. Due to the diversity of sensors the next and one of the most important steps in the processing chain is the harmonization of data.

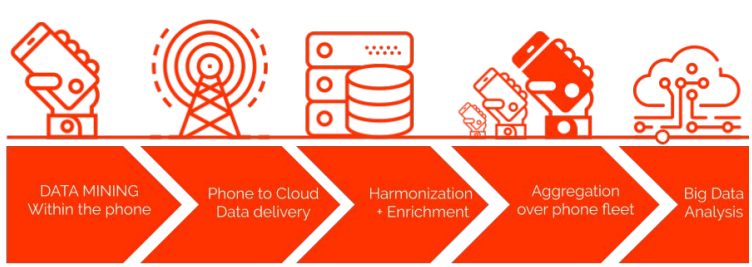

Fig. 1. Data Processing Chain for Big Data Aggregation

The heterogeneous nature of sensors and the variety of aggregation methods need to be lumped and standardised. Informations transmitted from the device into the cloud are often compressed and thinned out to save bandwidth. Enriching data extracted from devices ensures equality of measurements even from different sensors. Subsequently in the processing stage follows the aggregation of data over a whole fleet of devices. As one individual may not observe certain phenomenons, a whole fleet will as it increases the number of observed samples. The last step is the analysis of the aggregated data using Big Data techniques and methods, that extract information and knowledge to serve as basis for existing and novel cross-sectorial services.

In this paper we explore the potential of a holistic decentralised digital marketplace architecture for Big Data aggregation and exchange enabling service providers to ramp up large scale applications. Hereby, the CCDF data model, that harmonizes and generates brand independent datasets, is defined.

II. RELATED WORK

With the wide application of wireless sensor networks (WSNs), secure data sharing in networks is becoming a hot research topic and attracting more and more attention. A huge challenge is securely transmitting the data from the source node to the sink node. Except for eavesdropping the information stored in the packages, the adversary may also attempt to analyze the contextual information of the network to locate the source node.

Current devices-to-cloud systems rely on proprietary protocols and do not allow unrestricted access to data. One first approach of a standardized sensor data access is provided by World Wide Web Consortium's (W3C) Data standard. The standard, which is still under development, focuses essentially on the interfaces and an evaluation of the applicability of the protocol for automotive industry. In addition, the International Standardization Organization (ISO) enforces standardized data access interfaces using a neutral server instance with the focus on diagnosis services.

III. MARKETPLACE ECOSYSTEM & ARCHITECTURE

The architecture of the proposed Decentralised Big Data Marketplace shown in Figure 2 can be split into different components ordered around the central data exchange. Aggregated data are enriched using OEM internal knowledge to transform proprietary datasets into generic datasets. The OEM data refinery harmonizes proprietary information and brings them into CCDF standardized format. This step is one of the most important to provide a mutual understanding of the data for all entities involved in the data processing chain. Due to the great variety of sensors no unified, open and non-proprietary data model exists. Therefore the **Cross-Curated Data Framework** as new data model is introduced in the later sections. Harmonized and enriched data is stored encrypted inside the user device's personal cloud storage vault, depicted as *Data in the Cloud* in Fig. 2. The cloud storage keeps data in separate and individual vaults for each user to enforce maximum privacy protection by design. Thereby, the *Data in the Cloud* module follows the concept of citizen empowerment: the owner and creator of data stays in full control. Data is only offered and shared on the Big Data Marketplace, when the permission is actively given.

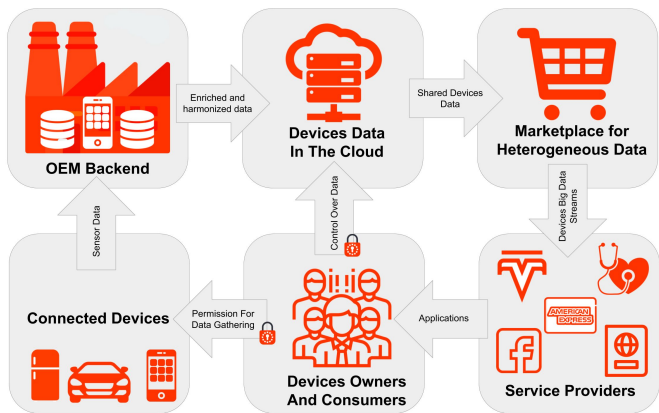

Fig. 2. Proposed Big Data Marketplace Architecture and Data Flow

The marketplace plays a central role in the presented architecture. It serves as an exchange platform for data and fulfills the tasks of a data accumulator and intermediary between owners and service providers. Here, data from different sources merges together. On the supply side, owners offer data. Incoming data needs to be preprocessed and indexed to allow searching and an availability analysis.

On the demand-side, service providers request data. They are able to search for available data and filter by region, age or quality. The Marketplace manages all offers and requests, collects the data from the individual user's cloud storage and combines it to Big Data from a whole fleet.

In the last step of the processing chain, service providers build innovative applications based on standardized and reliable datasets. These applications and services are offered in return to the users and other consumers.

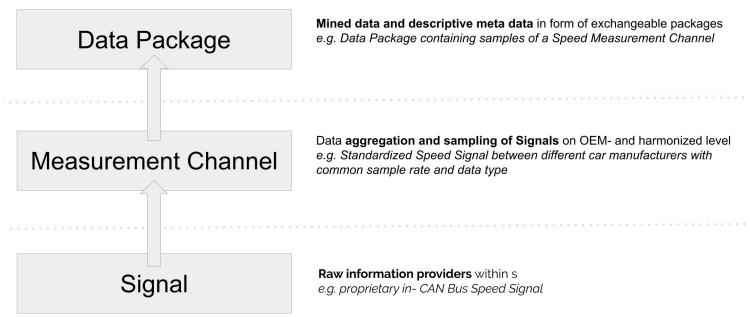

Fig. 3. Layered Data Model Architecture

IV. CROSS-CURATED DATA FRAMEWORK (CCDF)

A uniformised and efficient data structuration, harmonizing proprietary data as well as removing brand-specific information is a key success factor for the industrialisation then large adoption of the proposed Decentralized Big Data Marketplace. As there are no standardized, non-proprietary data models, the Cross-Curated Data Framework has been designed to be implemented within the a Big Data Refinery.

The CROSS-CURATED DATA FRAMEWORK relies on three interconnected layers (cf. Fig. 3). On the lowest layer, signals extracted from devices are aggregated. These signals are specific to both manufacturers and device and thereby proprietary. Their origin may be any source within the device, like. a speed signal captured from the accelerometer or On-Board Diagnostics (OBD) bus. The middle layer consists of Refining and Measurement Channels. Evoked data refinery define the common ground between signals and information originating from various industrialized devices. Referred layer enable the harmonization of proprietary information into a standardized format by removing brand specificities. On the top level, aggregated data is stored inside CROSS-CURATED Data Packages, A.K.A Dynamical knowledge bases.

These packages are exchangeable digital assets that can be transferred from devices via the respective OEM backend into the cloud storage to another devices.

A. Signals

Signals are the main material fueling Big Data processing chain. They are extracted by sensors, which work as the perceptive organs of devices and information systems. Sensors digitize the observable world by measuring physical, chemical and radio phenomenons translating them into binaries representations. More recently it has been proposed a subtle differentiation between digitization and its next step, datafication, i.e. putting a phenomenon in a quantified format so that it can be tabulated and analyzed. The fundamental difference is that digitization enables analog information to be transferred and stored in a more convenient digital format while datafication aims at organizing digitized version of analog signals in order to generate insights that would have not been inferred while signals were in their original form. Digital sensors enable digitization while connection let's data be aggregated and, thus, permits datafication. According to Gartner, by 2020 there will be 26 billion devices on earth, more than 3 devices on average per person. The pervasive presence of a variety of objects (including mobile phones, sensors, Radio-Frequency Identification - RFID - tags, actuators), which are able to interact with each other and cooperate with their neighbors to reach common goals, goes under the name of the Internet of Things, IoT. This increasing availability of sensor-enabled, connected devices is equipping companies with extensive information assets from which it is possible to create new business models, improve business processes and reduce costs and risks. In other words, IoT is one of the most promising fuels of Big Data expansion.

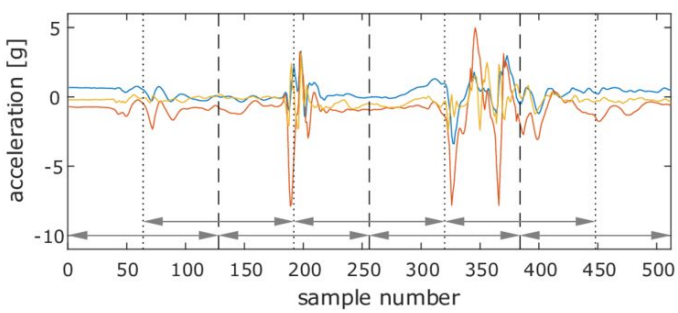

Fig. 4. Accelerometer signal (x, y, z) for incremental speed measurement

One major problem, which occurs at this stage and which the CCDF takes care of, is the great heterogeneity of sensor landscape. The variety comes from the large number of different OEMs and their suppliers. Different companies use different sensors for the same observations. Within the same manufacturer's production line, sensors can vary between models and sometimes between cars of the same model due to changes of supplier. This infers gaps in tests' rate, resolution and biases. Additionally, sensor configurations vary with the onboard Advanced Distribution Management Systems (ADMS) like parking assistant Therefore, two vehicles of the same brand would not provide the same set of signals.

Using the CROSS-CURATED DATA FRAMEWORK (CCDF) to extract signal extracted solves this problem by transforming signals into quantitative and qualitative information measured and classified by type in time. Processed information is then stored in distributed specialised containers and described in a standardized manner.

Current CCDF is based on BRAINCITIES' HR Dynamical Knowledge base build upon +3 000 000 heterogeneous signals and information types identified, categorised, measured and classified. They form the basis for signal measurements on the next layer, and are able to fuel any kind of information system with pondered and reliable information.

B. Measurement Channels

Many researches have been conducted in big data measurement processing. Despite the fact these studies have differing trends and results, they are fundamental to understand the whole industry. *Hal Varian and Peter Lyman from the University of California Berkeley are among the first researchers to study the field of measurement of volume of data produced, stored, and transmitted.*

As part of their "How much information?" project that ran from 2000 to 2003, they assessed that 5 exabytes of new data were stocked universally in 2002 and that 92% of the new information was stored on magnetic media, mostly in hard disks. It is an initial exhaustive exploration to quantify, in computer storage terms, the total amount of new and original information created in the world by the year and stored in four physical media: paper, film, optical (CDs and DVDs), and magnetics.

Since 2007 the study firm IDC to produce an annual series of reports on the "Digital Universe" to measure the amount of digital information created and recorded each year. The scientists appraised that in 2007 all the on hard drives, tapes, CDs, DVDs, and memory in the market equaled 264 exabytes. The basic methodological approach of the IDC in measuring of the Digital Universe was described as follows: Progress a prediction for the installed base of devices or applications that could capture or create digital information. Guesstimate how many units of information (files, images, songs, minutes of video, phone calls and so on) were created per year. Modified these units to megabytes Evaluate the number of times a unit of information is replicated. IDC research is based on more than 40 devices. Performs digital data calculation across the world and nearly 90 countries. IDC estimates that in 2016, the amount of digital information being created, captured, and replicated passed over 9.3ZB.

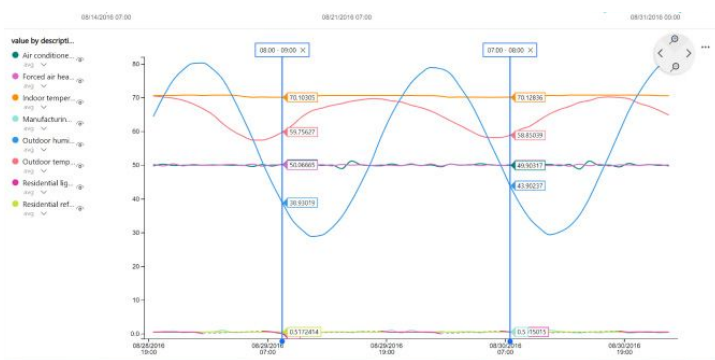

Fig. 5. Time Series Measurement in IoT solutions in near real time

Measurement Channels describe how data is extracted from signals. In devices sensors are polled and evaluated hundreds to thousands of times per second. With in average 20 to 100 different sensors installed, this produces an humongous amount of data. Storage limitation and data transfer capacity, make in-devices data preprocessing necessary. The obtained compressed data are then intégrated within the CROSS-CURATED DATA FRAMEWORK (CCDF) Measurement Channels.

Focus on Time Series Measurement
TSMs aggregate signal values over time. Hereby, the underlying signal's sample rate can be reduced by either keeping only one sample per interval (downsampling) or averaging over several samples. Fig. 4 shows an excerpt from a typically engine speed signal (sample rate 100 *Hz*) that is downsampled to meet the 1 *Hz* sample rate of the engine speed TSM.

DATAWALLET ARCHITECTURE

The Datawallet System mainly deals with storage and security of data. It combine a cloud storage solution built with PHP for the storage services and custom NodeJS scripts for automations like user management, file transfers, auto deployments, payment integrations, managing share permissions etc.

Types of groups

Each group has its own private server meaning every group has its own instance of the cloud storage solution running on a dedicated server. Every group have their own members and these members have a space inside the particular server. This space is called the Datawallet. So the terms Users and Datawallets can be used interchangeably. So the network as a whole is called the Datawallet System or the Datawallet Network, the groups inside this network represents 1 institution and every group has its own users/members/datawallets.

Every group is identified by a unique id called the Group Public Key. Format is dw-xxxxxx

The Datawallet System consists of two types of groups:

- **Federation type group**
  These groups are real world institutions like Swiss bank, Porsche etc. Every member/datawallet inside a Federation represents a sub-institution. For example if the a group is represents Swiss bank every datawallet inside Swiss Bank represents a bank that comes under Swiss Bank i.e a branch of Swiss Bank.

  Categories of federations:
  a. Banking
  b. Healthcare
  c. Automotive
  d. Mobility
  e. Hospitality
  f. Human Rights
  g. Energy
  h. Agriculture
  i. Public Institutions

Every group of type federation will lie in one of the above mentioned categories.

- Community type group
  These groups contain real world users. For example if a community group represents Paris then every datawallet inside it represents a user of paris.

  Categories of communities
  a. Cities
  b. Objects
  c. Musicians
  d. Real Madrid
  e. Bots etc

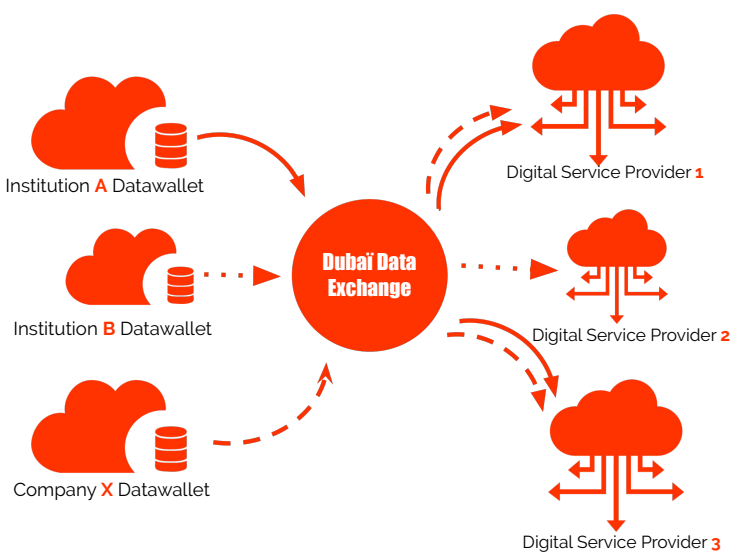

Fig. 3. Dubai Data Exchange Manages Data Flows Between Private Data Storage Vaults and Service Providers

Types of users

Every user is identified by a unique id called the User Public Key.
Every user also has a password called the User Private Key which is used for logging inside the storage space and the user dashboard.

1. **Admin**
   a. Creates new template folders for all users
   b. Manages a group with public key = dw-unattributed. This group is basically a community that consists of users who are not associated with any specific community.
   c. Does not have its own datawallet

2. **Federation operator**
   a. Purchases credits
   b. Creates new federation users for his own federation
   c. Modifies group description like group name, group info
   d. Also functions like a normal federation user
   e. Accesses files present in his datawallet

3. **Federation user**
   a. Purchases credits
   b. Sends folder access request to other federation and community users
   c. Sends files to other federation and community users upon request being accepted by them
   d. Accesses files present in his datawallet

4. **Community operator**
   a. Purchases credits
   b. Created new community users for his own community
   c. Generated QR codes to be able to add new members to the community easily
   d. Modifies group description like group name, group info
   e. Also functions like a normal community user
   f. Accesses files present in his datawallet

5. **Community user**
   a. Accepts / rejects requests sent by federation users
   b. Accesses files present in his datawallet

Types of credits

1. Private credits
   a. used to create new group members by the group operator.
   b. 1 credit = 1 new user

2. Target federation credits
   a. used to send folder access request to other federation members.
   b. 1 Target federation credit = being able to upload files to 1 folder of another federation user(single)

3. Target community credits
   a. used to send folder access request to other community members.
   b. 1 Target community credit = being able to upload files to 1 folder of another community user(single)

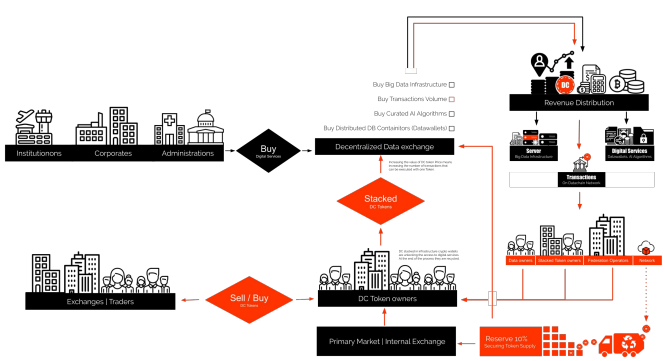

Fig. 4. Business Model - Transaction-based model with a native token integrated within the exchange system to incentivize data sharing

5. **Community user**
    a. Accepts / rejects requests sent by federation users
    b. Accesses files present in his datawallet

Types of credits

1. Private credits
    a. used to create new group members by the group operator.
    b. 1 credit = 1 new user

2. Target federation credits
    a. used to send folder access request to other federation members.
    b. 1 Target federation credit = being able to upload files to 1 folder of another federation user(single)

3. Target community credits
    a. used to send folder access request to other community members.
    b. 1 Target community credit = being able to upload files to 1 folder of another community user(single)

V. THE BIG DATA MARKETPLACE

The Marketplace is a mediator platform that offers access to standardized information, which is represented using the CCDF. It ensures the flow of data from the owners' private cloud storage to the service providers' backends to enable the creation of new services and applications.

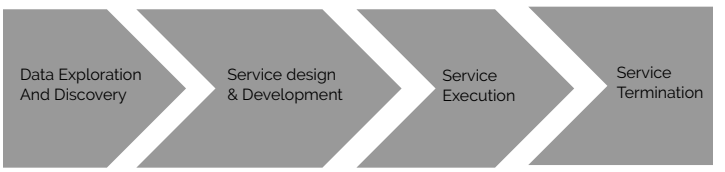

Fig. 6. Heterogeneous Big Data Service Life-cycle Phases

The marketplace includes key features to cover the complete life-cycle of Big Data services as shown in Fig. 6. All features are exposed via two different interfaces: one REST (REpresentational State Transfer) interface, specified in a machine-readable format for seamless integration with services providers' existing infrastructures, and one responsive cross-platform front-end for human-friendly interaction.

DUBAI DECENTRALIZED DATA EXCHANGE
A Practical Use of BRAINCITIES' Decentralized Data Analytics Infrastructure.

The Marketplace is an online store that allows a federation members to make and sell a dataset as a product. The dataset is created from the Datawallets that the Federation Member has been granted permission to. These permissions are read type permissions that the Federation Members request from Datawallet owners. Read type permissions can be requested from Datawallet owners of various private and public communities.

Communities: It is a group of people who have a storage space known as Datawallet.

Federations: It is a group of people that can have access(read/write) to the Datawallets that belong to a community but only after the Datawallet owner grants permission.

For the Dubai Smart City Federation, its members can be:

- Telecommunication industry ecosystem
- Hospitals/Healthcare industry ecosystem
- Transport industry ecosystem
- Automotive industry ecosystem

For Dubai Federation, its members can be:

- Ministry of Happiness
- Ministry of Education
- Ministry of foreign affairs

For the Dubai Smart City Community, its members are the people of Dubai who have a personal storage space known as Datawallet.

Working of Marketplace

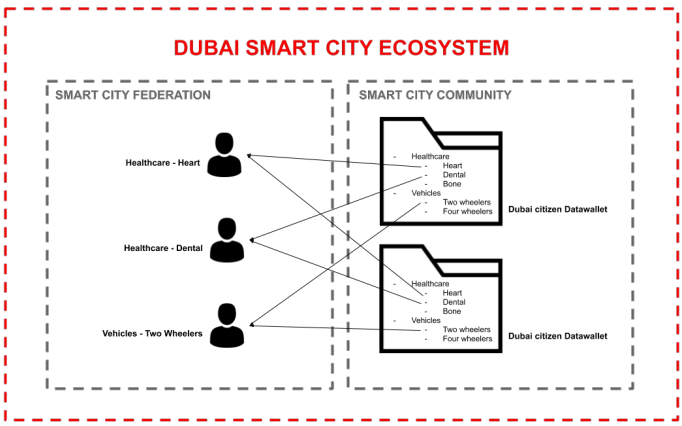

Fig.7: Datawallet structure and permissions

1. *Permission request*

A federation member requests for a read type permission for some users of a community. Example: Dubai Smart City Healthcare - Heart requests for read type permission of the Heath/Heart folder of all the users of the Dubai Smart City Community.

2. *Permission acceptance*

The Datawallet owner of the target community accepts the permission request thereby allowing the Federation member to have access to his/her particular folder the personal Datawallet.

Example: A Dubai Smart City citizen accepts the read type request of Dubai Smart City Healthcare thus giving the read permission of his/her Healthcare folder.

3. *Product creation*

The Federation member starts by creating a marketplace that can be accessed by a fully custom domain name of his/her choice. After this, new dataset/products can be added to his/her marketplace. These datasets/products have the following attributes that help in their creation:

- target folder
- update frequency
- dataset type: json,zip
- price etc

1. *Dataset caching*

The dataset is created by extracting information from the Datawallets of those community members who have granted the read access permission. Datasets are refreshed according to the chosen time interval.

2. *Product access*

After the dataset has been purchased by a customer, it can be either is downloaded as a zip file or it can be accessed via an API key in JSON format.

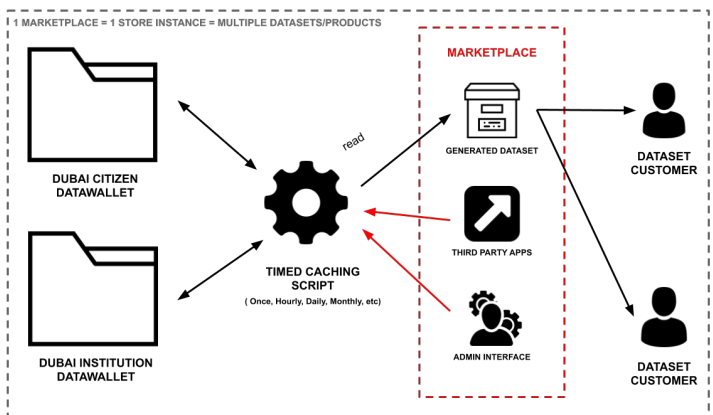

Fig.8: Caching process and Marketplace structure

B. Service Design and Development

For the initialization of data flows, access agreements between service providers and data owners are set up through smart contracts and managed by the marketplace. The marketplace integrates a message broker as part of its internal architecture as shown in Fig. 8. This message broker handles the routing of all data flows according to the agreements from the device owners' Datawallet to the service providers.

C. Service Execution

During run-time of the service, the marketplace transfers data from the device owners' Datawallet to the service providers according to the arranged agreements. The following data retrieval approaches are supported:

Pull Mode: In pull mode, service providers perform queries to receive data from the marketplace.The Marketplace validates, whether new data is available, retrieves and accumulates data from all cloud storage vaults, which is delivered afterwards. This approach features the data retrieval of one whole device-fleet in one request.

Push Mode: In the alternative push mode, the cloud storage forwards incoming CCDF Data Packages via the marketplace's message broker to the service providers according to access agreements. This approach minimizes the delay between observation and processing and enables highly fresh data; the accumulation for all individual devices to fleet data has to be performed by the service provider.

D. Service Termination

At any time, vehicle owners may decide to stop sharing their data, canceling the corresponding agreements. The data flow from the corresponding Datawallet is stopped. On the other side, service providers can as well end their data subscriptions, where data flow from all sources is terminated and unregistered from the their referred ecosystem.

VII. CONCLUSION

In this paper a concept for an holistic architecture for Heterogeneous big data aggregation is presented. The proposed system enables providers of services and applications to access devices and information systems data via a single point of access - the Decentralised Data Marketplace. The marketplace handles devices, individuals, information systems clusters accumulation and contract handling with entities/devices/information systems owners while preserving privacy rights of all data stakeholders. The proposed unified data format CCDF solves the tasks of harmonizing proprietary and non proprietary sensor data and information. CCDF allows different levels of quality as well as different data aggregation types in form of time-series and histograms.

As shown at the Bank examples the proposed Decentralised Data Marketplace enables exploitation of business related big data for innovative and cross-sectorial applications and services.

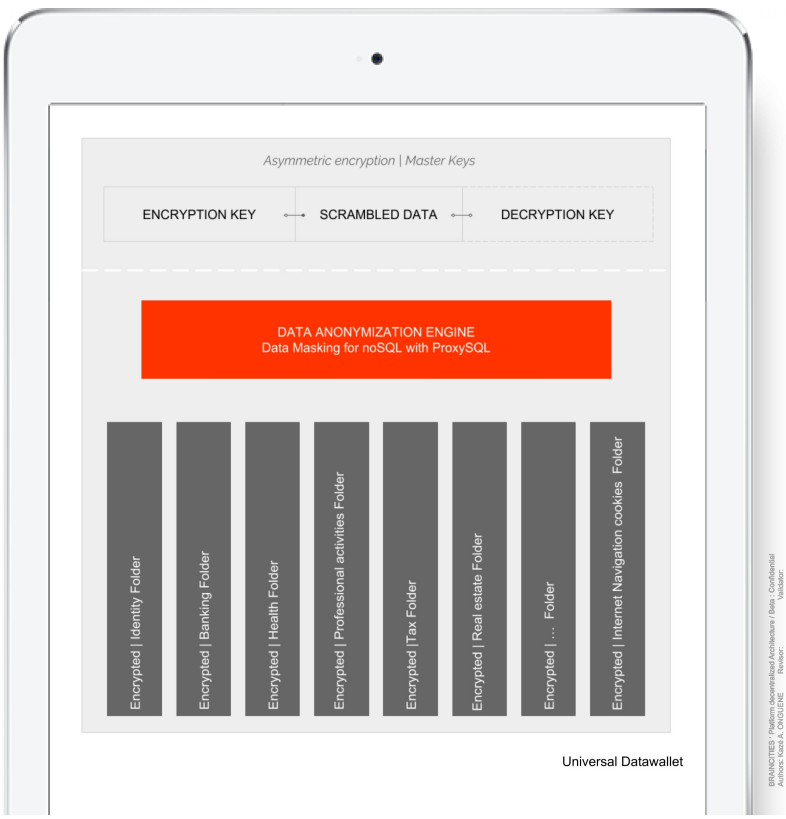

Fig. 9. Datawallet the Distributed data analytics infrastructure central unit

BRAINCITIES

DUBAÏ

Emirates Towers
Sheikh Zayed Road
Dubai, UAE - Po Box 31303
International. +447 413 341 581

PARIS

21 Boulevard Haussmann
75009 Paris France
phone +33 156 036 752