Practical Experience with Smart Cities Platform Design

Ang Loon Chan
Institute for Infocomm Research, Singapore
chanal@i2r.a-star.edu.sg

Gim Guan Chua
Institute for Infocomm Research, Singapore
ggchua@i2r.a-star.edu.sg

Desmond Zhen Liang Chua
Institute for Infocomm Research, Singapore
desmond-chua@i2r.a-star.edu.sg

Shuqiao Guo
Institute for Infocomm Research, Singapore
guosq@i2r.a-star.edu.sg

Paul Min Chim Lim
Institute for Infocomm Research, Singapore
limmc@i2r.a-star.edu.sg

Mun Thye Mak
Institute for Infocomm Research, Singapore
mtmak@i2r.a-star.edu.sg

Wee Siong Ng
Institute for Infocomm Research, Singapore
wsng@i2r.a-star.edu.sg

***Abstract*— A successful smart nation is one which uses information and Internet of Things (IoT) technology in a seamless integrated form to enhance transport, healthcare and other public services to improve the quality of life for citizens. In this paper we present our practical experience with smart cities platform design: A*DAX, a powerful framework for extracting insight from heterogeneous, real-time and complex data sets. In fact, A*DAX has been successfully deployed at the multi-agency Jurong Lake District ("JLD") Smart City Test Bed in Singapore, with integration to a shared sensor network of close to 1,000 urban sensors and video analytics of more than 25 video sensors. The vision for JLD is to be a leading model for developing a mixed-used urban area that is sustainable, smart and connected. Through the A*DAX data exchange portal and repository, test bed users will be able to access real-time and historical environmental data collected by the sensors, as well as video analytics from the camera streams and benefit from a variety of Smart Nation solutions providing real-time sensing and analytics.**

Keywords—A*DAX, Smart Cities, IoT, data analytics and exchange, data platform

I. Introduction

In a smart nation or city, large amount of urban data are routinely collected by both public and private sectors in increasing volume, rate and variety [3]. Analysis and central management of such data can lead to a big-picture view and a better understanding of how to meet the challenges in an urban setting. There is a need for a data exchange, data fusion and sense-making platform with capabilities to integrate, manage and analyze such collected data.

A*STAR Data Analytics Exchange ("A*DAX") [1,2,3] is a scalable platform built upon open architecture and is designed to facilitate the secure management and analytics of urban data. A*DAX can be deployed as the data aggregation portal and data analytics platform, with capabilities to collect and manage urban data, so as to enable machine learning, base-lining and trend analysis. With the appropriate data fusion and analytics, A*DAX can yield valuable and actionable insights, which can be tapped to enable the development of data-driven tools and applications. In fact, A*DAX has been successfully employed at the Jurong Lake District ("JLD") Smart City Test-Bed [8], with built-in backend integration that ingests and manages huge quantities of sensor and video data.

Many innovative ideas and applications are designed and deployed in JLD. They are broadly divided into three categories namely (i) urban mobility, (ii) sustainability and (iii) improving sensing and situational awareness. For example, a smart queue monitoring system that leverages on the power of advanced video sensing to determine length and flow of queues in real-time; smart bin technologies to automatically determine the cleanliness of public areas; multi-modal positional solutions to mitigate issues related to global positioning system (GPS) navigation in urban canyons situation; illegal parking detection through video sensing and so on. These applications require large datasets, complex data analysis and powerful computational resources for data integration, analytics and visualization.

In this paper, we will describe our experience with smart cities platform design which is a scalable data exchange platform built upon open standards and architecture to bring better situational awareness through data collection, efficient sharing and making sense of collected data through machine learning and data analytics.

II. Architecture

This section describes the architecture of the platform and how data flowing through the system are processed, analyzed and turned into actionable insights, alerts, visualizations and dashboards. Figure 1 shows the architecture of platform where the modular components work together to form an integrated and complete solution stack. Yet at the same time, it is flexible enough to integrate with and obtain data from external sensor gateways or data repositories which support open standards for data exchange.

a) Data Sources: These are not part of the platform, but rather external data sources with data that needs to be ingested into the platform for integration and analysis purposes. Assuming data can be pulled or pushed to the platform through open standards like FTP, MQTT or RESTful

APIs, our data adaptors can ingest it and share it with other users using a streamlined RESTful API for usage, download or analytics. It is assumed that the sensor data and metadata (cleaned-up by the sensor gateway) are in open standards like CSV, TSV or JSON format.

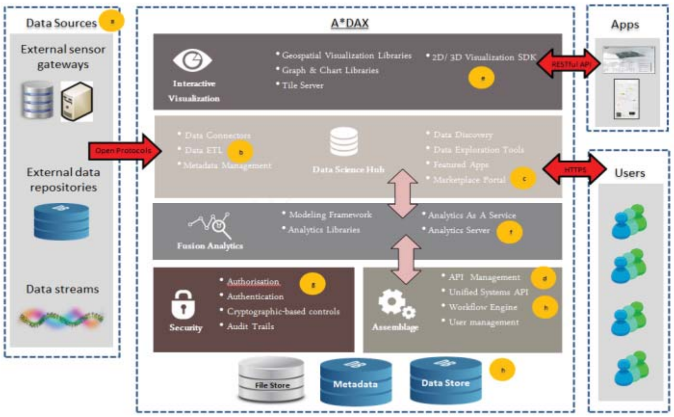


Figure 1: Adaptable and Unified Architecture

b) Data Extraction, Transformation and Ingestion: The platform has a Data ETL (extract, transform and load) framework that manages data flows, extracts data, transforms data and ingests it into metadata and data stores for later integration and analysis. This adaptable architecture allows custom adaptors to be developed where necessary.

c) Marketplace Portal: This is a one-stop web portal for authorized users to discover, explore, download and manage the data stored in the platform.

d) Data APIs: Data ingested in the platform are available for access by authorized applications and users via secure RESTful APIs. The API Service verifies with the API Management Server for valid access tokens and permissions before granting access to resources. The Data APIs are used by developers to develop data-driven applications or custom dashboards.

e) Reports & Visualization: The platform provides a Visualization SDK for the development of data trend charts and data visualization, which may require fusing (or at least co-displaying) different attributes and datasets.

f) Data Analytics Tools: The platform provides data exploration and data analytics tools in the form of Data Preview within the Data Marketplace Portal. Here, users can preview ingested data in tabular or graph formats. If the data contains geospatial information, it can also be previewed in map format. These tools are available out-of-the-box without the need for customization and facilitate exploratory analysis of data. The platform also provides libraries that make it easy to develop cross-filter charts that allow users to compare trends and analyze correlations of attributes within a dataset.

g) Security: Important system events and errors are logged in Audit Trails while access to resources is only granted to authorized users and applications.

h) Assemblage: This is the part of the architecture that forms a unified command, control and operation hub for higher-level component blocks within the platform. It provides API endpoints to facilitate easy integration with legacy and backend systems, and also provides scalable compute & storage support for data science activities.

III. Capabilities & Functions

The following sections provide a list of the key components, capabilities and functions provided to enable the secure management, analytics and exchange of data, so as to facilitate the extraction of actionable insights.

A. Data Science Hub

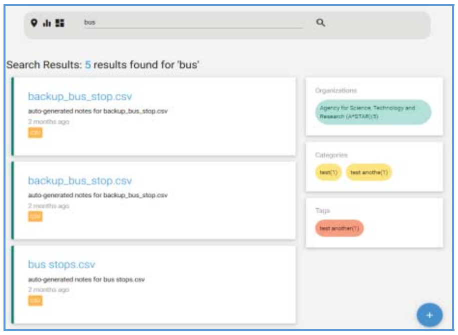
The platform is designed to connect data silos and to make data useful. It consists of various components, of which the Data Science Hub is the database- and language-agnostic data platform for the organizing, managing, sharing, collaborating and processing of data. It supports high-throughput and distributed data access, and allows users to programmatically access a wide variety and huge volume of data via a set of open standards Data APIs.

This section lists the components for providing the components and capabilities of the Data Science Hub:

a) Marketplace Portal

This is a one-stop web portal for authorized users to discover, explore, download and manage the data stored in the platform. It has the following functions:

i. Data Discovery


bus

Search Results: 5 results found for 'bus'

backup_bus_stop.csv
auto-generated notes for backup_bus_stop.csv
2 months ago
CSV

backup_bus_stop.csv
auto-generated notes for backup_bus_stop.csv
2 months ago
CSV

bus stops.csv
auto-generated notes for bus stops.csv
2 months ago
CSV

Organizations
Agency for Science, Technology and Research (A*STAR)(5)

Categories
test(1) test anothe(1)

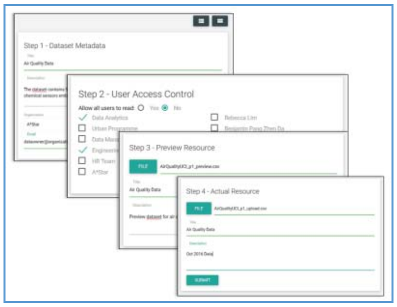
Tags
test another(1)

Figure 2: Keywork Search

The built-in search engine in the portal allows users to discover the correct data required for a specific purpose, since un-searchable data in the data warehouse is unusable, useless and a waste of space. Figure 2 shows a typical search result screen after a keyword search. From this screen, the user will be shown other categories of related data which may be of interest and he can utilize those links to expand his search for data.

ii. Metadata Management

The metadata management system is capable of handling both standard metadata fields (like title, update date-time stamp) and domain-specific fields (like sensor type, frequency) without compromising on the ability for the search engine to search across all the fields (both standard and domain-specific). Figure 3 below shows screen captures from the metadata management workflow.



Step 1 - Dataset Metadata

Air Quality Data

Step 2 - User Access Control

Allow all users to read: ○ Yes ● No

✓ Data Analytics

□ Urban Programme

□ Data Mas

✓ Engineerin

□ HR Team

□ A*Star

□ Rebecca Lim

□ Benjamin Pang Zhen Da

Step 3 - Preview Resource

FILE AirQualityIJC3_p1_preview.csv

Air Quality Data

Step 4 - Actual Resource

FILE AirQualityIJC3_p1_upload.csv

Air Quality Data

Oct 2016 Data

SUBMIT

Figure 3: Workflow for Metadata Entry

iii. Data Catalogue

The Data Catalogue lists datasets which are organized according to data categories, subject tags and data sources.

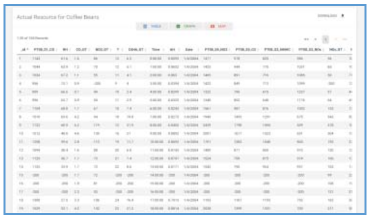
iv. Data exchange

The portal supports secure exchange of data between authorized users. Data can be downloaded via secure HTTPS protocol in CSV format. Alternatively, data can also be accessed via RESTful APIs.

v. RESTful Data API

The platform provides RESTful APIs with responses in JSON format for authorized developers to query, access, search, download and use the datasets in a secure manner. Both REST and JSON are open and common standards for web-services which are easy to use.

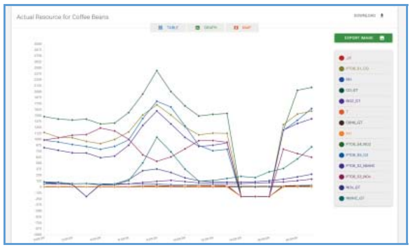
vi. Visual Analysis Tools



Actual Resource for Coffee Beans

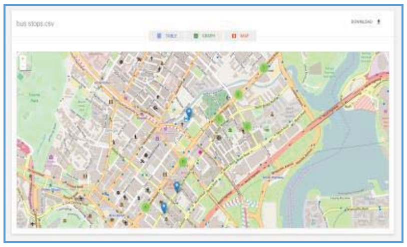
Figure 4: Table View

The platform provides data exploration and data analytics tools in the form of Data Preview within the Marketplace Portal. The preview tools provide visual analysis capabilities by allowing users to visualize structured data through tabular (Figure 4), map (Figure 6) or graphical (Figure 5) views. It is highly interactive and is primarily used by technically-savvy users like data analysts and researchers during exploratory analysis to find anomaly trends, identify events' correlations, patterns, and business conditions in data with respect to observations.



Actual Resource for Coffee Beans

Figure 5: Graph View



bus stops.csv

Figure 6: Map View

b) Data ETL

A*DAX can ingest, store and manage heterogeneous data. The process of extracting, transforming and loading (ETL) the incoming external data into an internal data store is to facilitate data analytics. The platform supports parallel and scalable data routing based on directed graphs. It also allows user to define customized transformation functions and system mediation logic.

c) Timestore Add-Ons

For the purposes of visualisation, there is often a need to retrieve time-series data. When dealing with large time scales, individual records contribute less information than their aggregate, even for trend analysis. As such, the Timestore Add-on was designed to provide a multi-scale view of time-series data, which is especially relevant for sensor data.

Timestore is an add-on to any existing datastore that has time-series data. Timestore provides two features for trend analysis: aggregate caching and time-scale regulation. From these, additional insights can be developed to speed up other related computation.

i. Aggregate Caching

At large time scales, various descriptive statistics (e.g. mean, variance, skew, and kurtosis) are more useful for analysis than raw records. The traditional way of doing this is to compute the statistics required for records within a time slice, before finally displaying the results visually. This is inefficient because a lot of data being retrieved gets discarded almost immediately after computation. If there is a drill-down step, even more of the data are discarded, wasting the resources it took to retrieve them in the first place.

Timestore addresses this inefficiency by providing tables that pre-compute the aggregated quantities like $\sum x$, $\sum x^2$, $\sum x^3$, min(x), max(x), mode(x), and useful numbers like the number of samples (n) for a given time slice size (in integer seconds). From such aggregated values, various standard descriptive statistics (e.g. mean, variance, skewness and kurtosis) may be computed and returned via the Timestore API.

The strategy of aggregate caching can be further extended to compute other types of statistics useful to different types of analytics.

ii. Time-Scale Regulation

Timestore simplifies retrieval of the aggregated values through enforcing fixed duration periods per sampling block. This means that each record that occurs in the Timestore tables provides aggregated information over the same duration over time. For example, if the fixed duration period per sampling block is at 3,600 seconds (1 hour), then each record in the Timestore represents aggregated information over each hour.

This time-scale regulation means that the retrieval of statistics given the duration of the time slice is constant regardless of the in-coming data rate. So, if the first hour has a spike of data (say 500k samples), and the next hour has only a small amount of data (say 1k samples), the time needed to retrieve the mean for each hour is the same.

d) Trend Functions

The platform provides a few common curve-fitting functions to aid in visualizing time-series data. The Linear Regression method can be used to get quickly compute the trend line.

Another three methods are commonly used to smooth out "noisy" time-series data:

- Simple Moving Average (SMA)
- Exponential Smoothing (ES)
- Double Exponential Smoothing (DES)

SMA is the simplest, but it takes into account only a fixed window of historical data. ES is computationally simpler and only requires the most recent forecast values to be kept. DES is used when there is a trend in the data. It is recommended to use forecasting packages (such as those found in R) to get the appropriate parameters beyond the above common tools.

e) Fusion Analytics Framework

The adoption of a common and open framework facilitates a unified and sustainable approach to data analytics. The adaptable and open architecture of the platform is designed to facilitate machine learning and data analytics through the Fusion Analytics Framework. This framework enables rapid development of data-driven web and client applications through extraction of the complex insights behind machine learning and analytical functions.

When large volumes of diverse data are collected by an organization, it requires the power of Big Data Analytics to make sense out of them. This will inevitably use complex analytics functions and primitives developed for domain-specific use cases. Without a suitable framework to streamline the processes of Big Data Analytics, the developed analytics tends to end up as one-off development projects and not re-usable.

The Fusion Analytics Framework is an open and extensible Big Data Analytics framework that enables rapid development of data-driven solutions. Its key value propositions are as follows:

i. Allows developer to focus on analysis instead of re-learning methodologies;

ii. Ease-of-use for Big Data Analytics solution developers; and

iii. Sustainable and unified approach to Big Data Analytics for continuous insight generation.

The framework supports the development of Data Analytics Toolkits that can be used to build Big Data-driven applications or solutions. The unified and reproducible approach to Big Data Solutions leads to a sustainable infrastructure and development framework that enables the customer to fully leverage on Big Data Analytics to drive future innovations.

B. Interactive Visualization

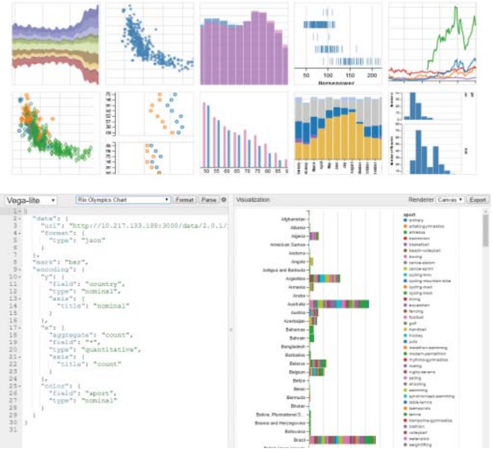
A*DAX provides a set of open-standards data visualization tools that allows users to create effective data visualization programmatically. It is intuitive to deploy with minimal coding requirement and is designed to improve responsiveness and address scalability issues. The key component necessary is the Data Visualization SDK that comes with the platform.

a) Data Visualization SDK

The Visualization SDK provides the standards by which a visual display of the most important indicators and performance measures can be consolidated and arranged on a single screen, so the information can be monitored at a glance.

Most importantly, the data, information and analytic results to be displayed are pulled from the platform in a secure manner using standard RESTful APIs or real-time queries through web-sockets. The platform is designed to prohibit direct access to the database so as to enforce role-based access control, to ensure data security and data integrity.

As the output is in standard HTML5, it can be viewed directly in most modern web browsers or embedded into other third-party dashboard systems. The Visualization SDK is designed for the development and deployment of customized data visualizations and insights exploration, where the requirements involve the fusing of (or at least co-displaying) vastly different datasets from vastly different domains.



```
{
  "data": {
    "url": "http://10.217.133.188:3000/data/2.0.1/
    "format": {
      "type": "json"
    }
  },
  "mark": "bar",
  "encoding": {
    "y": {
      "field": "country",
      "type": "nominal",
      "axis": {
        "title": "nominal"
      }
    },
    "x": {
      "aggregate": "count",
      "field": "*",
      "type": "quantitative",
      "axis": {
        "title": "count"
      }
    },
    "color": {
      "field": "sport",
      "type": "nominal"
    }
  }
}
```

Figure 6: Code Playground

C. *Assemblage*

This forms a unified command, control and operation hub for higher-level component blocks within the platform. It provides API endpoints to facilitate easy integration with legacy and backend systems, and also provides scalable computational & storage support for data science activities.

This section lists the components necessary for providing the capabilities of Assemblage:

- API Server and Management

The API Server is designed for a secure and controlled access to backend services and data. The API Server can be seen as 2 separate services: the Internal Systems API Services and External API Services.

The Internal Systems API Services are used by internal subsystems that require frequent access to back-end assets like the Data Reservoir. The External API Services are for external users, where more robust and stringent security control is needed. Only the External API Services will be described in this section.

The API Server allows for sharing of information and services in a consistent and secure manner. For example, service providers can access data and services from the platform to offer user apps and dashboards through data access APIs. A clear and comprehensive set of API standards and protocols are published for developers to interface with and push services/data to other externally developed applications and external users.

The lifecycle, usage and security of the APIs needs to be managed properly from a central location for ease-of-use and accountability.

To provide all these, the API Server and Management is powered by a scalable API life-cycle management framework which includes the processes of creating, managing, publishing and securing APIs. APIs' behavior can be monitored and analyzed in real-time though the available reports and alerts from the system.

- Data Reservoir

This is the Big Data Infrastructure which allows the platform to store and manage data of diverse varieties, huge volumes and high velocity. The valuable data is stored here to form a scalable Data Reservoir that is ready to be "piped-out" to the other components to be processed, transformed or analyzed. The Data Reservoir is sufficiently adaptable to be able to store data in the form of discrete files or relational data models.

Figure 7 below shows the architecture of the Data Reservoir with higher level abstraction software layers to allow internal processes and applications to access database resources in a secure and structured manner.

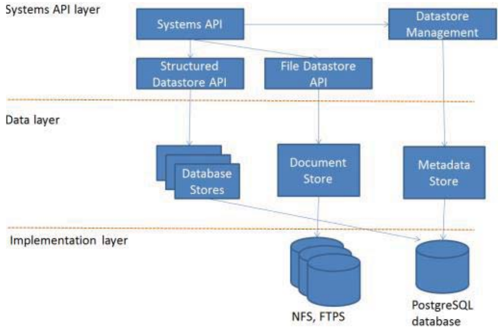


Figure 7: Data Reservoir & Systems API

- Unified Systems API

The Systems API provides the other components with a consistent method to access the different types of data from the Data Reservoir. The data is referenced by an unique identifier of the datastore. The backend data implementation is transparent to the end-user, so data in the same class (e.g. structured data or file) can be accessed by the user without having to deal with the complexity of each data type.

For example, the Structured Datastore API presents functions to manipulate structured data with SQL-like options. Likewise, the Filestore API presents common functions to manipulate files.

Multiple backend datastore implementations may be configured within the same class of data. For example, the structured datastore API can make use of two PostgreSQL instances to store different tables. In order to retrieve the data from the correct source, there exists a lookup table of datastore IDs (part of Metadata store) to allow the retrieval of the correct datastore instance and credentials in order to query the data.

IV. SECURITY

A. *Platform Security*

Security is an important aspect of any data platform. It is assumed that the platform administrator will ensure the data security of data in the static data repository. Authorization services will be provided by the platform administrator for users to access the data stored in the static data repository, and this is managed via Identity Management and Access Control components. The following safeguards are implemented as layered defenses to prevent unauthorized access or modification of the data:

a) Only authenticated and authorized users can access the data

b) Data and requests submitted to the API Server is encrypted through SSL/TLS.

c) Data and requests downloaded from the API Server is encrypted through SSL/TLS.

d) To protect information in transit, channel encryption will be implemented with 256-bit SSL for HTTPS, with up to SHA-2 signing capabilities and 2048-bit keys

e) Access and audit logs for API access to data are recorded (read, update)

f) Access control for API access to data streams and analytic functions is assumed to be provided by the Access Control Service

g) All APIs require valid API keys and token for authorization and authentication

B. *Platform Audit Trail*

To ensure user accountability and action traceability, the platform provides the means for user actions to be traced, by implementing the following audit trail and logging functions:

a) Audit logs for security events such as user authentication, API usage with invalid API token, API usage with API token (except for read operations), dataset downloads, dataset / meta-data updates

b) Access and error logs

c) For user-centric events at API Server such as Data and time of the event, user ID (if applicable), nature of event, action taken.

V. RELATED WORKS

There are numerous open data initiatives for Smart Cities. In United State, Data.gov [5], which is powered by the open sourced CKAN [6] data platform. It includes more than 190,000 searchable datasets with more than 100,000 geo-spatial datasets in finance, education, agriculture, healthcare and etc. The New York City Data Mine [7] provides many sets of public data produced and used by New York City government. Applications like map and chart crime statistics with the data from New York City Police Department and application to find public Wi-Fi in the neighborhood. The Open Data Catalogue of the City of Vancouver provides information about parks, parking, crime, schools and transportation data in CSV, XLS, KML and SHP formats. Sweden's opengov.se is led by private entity aiming to collect and share available public datasets in Sweden. The main focus of these initiative is to provide a directory service of mostly data files for public use, whereas A*DAX is a single integrated processing environment that supports structured and unstructured data, as well as ad-hoc and continuous query for the purpose of data exchange and analytics.

VI. CONCLUSION

In this paper, we introduced A*DAX, which is the latest developed version based on the previous works [1,2,3,4]. It is an open system and streamlined the data analytics processes. As an open system, many standards are supported by the platform's native components, particularly on security and support for geospatial data processing and sharing. Its capabilities and adaptable architecture will enable public agencies, private businesses and the citizens of a smart nation or city to make informed decisions and respond to dynamic conditions based on real-time sensing and data analytics. The system has been technically proved and deployed in real-world environment, i.e., The Jurong Lake District (JLD) Testbed Project [3].

ACKNOWLEDGMENT

The project was supported by the A*STAR SERC Grant No. 1524100032.

REFERENCES

[1] Amudha, N., Chua, G.G., Foo, E.S.K., Goh, S.T., Guo, S., Lim, P.M.C., Mak, M.T., Munshi, M.C.M., Ng, S.-K., Ng, W. S. and Wu, H. (2014). "A*DAX: A Platform for Cross-domain Data Linking, Sharing and Analytics", DASFAA 2014.

[2] Lim, P.M.C., Ng, S.-K., Ng, W.S., Wu H. and Quek, A.M.H. (2014) "A*DAX for Transport Data Management, Sharing And Analytics", ITS World Congress 2014.

[3] Lim, P.M.C., Ng, S.-K., Ng, W.S., A*DAX: Data Analytics Platform for Smart Cities. Smart City Expo World Congress 2014

[4] Yu, L., Yee, J.C.P., Ng, W.S., Lim, P.M.C. and Ng, S.-K. (2014) "SINGAPORE-IN-MOTION: More Data Lead to Better Understanding", ITS World Congress 2014.

[5] Data.gov., https://www.data.gov/

[6] Katarzyna, O. and Jarosław ., Open Data collection using mobile phones based on CKAN platform, FedCSIS, 2015

[7] NYC Opendata, https://opendata.cityofnewyork.us/

[8] Innovation for a Smart Nation, http://www.i2r.a-star.edu.sg/sites/default/files/online-kit/FINAL%20Astar%20Smart%20Nation_single%20page.pdf