A Case Study: Phuket City Data Platform

Aziz Nanthaamornphong
College of Computing
Prince of Songkla University
Phuket, Thailand
aziz.n@phuket.psu.ac.th

Jeffrey Holmes
College of Computing
Prince of Songkla University
Phuket, Thailand
jeffreyholmes@hotmail.com

Pracha Asawateera
Digital Economy Promotion Agency
Bangkok, Thailand
pracha.as@depa.or.th

***Abstract*—The Phuket 2020 Smart City initiative, with its two strategies - Smart Economy and Smart Living Community, aims at creating greater trust in the government through responsiveness, transparency, and providing opportunities for greater engagement between government and citizens. One of the prerequisites for the successful implementation of the project is a big data platform that integrates several existing data sources around Phuket into one central hub and provides information and new insights to public and private entities. To achieve that, it is not enough to simply collect as much information as possible. The various sources must be put in a unified context that is enhanced with publicly accessible data and processed using methods and tools that are capable of mining new insights. The Digital Economy Promotion Agency has identified various data sources and is preparing for the next phase of data integration into a big data platform called the Phuket City Data Platform. The objective of this paper is to present a framework and a high-level action plan on the integration of these data sources. The paper highlights potential business values in four key areas: tourism, safety, environment, and the digital industry.**

***Index Terms*—Smart city, big data, data platform.**

I. Introduction

The Phuket 2020 Smart City project is part of the Thai government's digital economy policy initiative to increase the competitiveness of the country by creating seven smart cities throughout Thailand. The Digital Economy Promotion Agency (DEPA) was created to execute the Smart City project. Phuket is set to be the first smart city in Thailand due to its numerous tourist attractions, outstanding natural resources, highly stable communication network, and strategic development towards becoming a technology business hub in the near future.

'Phuket Smart City Data Platform' involves seven elements: Smart Economy, Smart Environment, Smart Governance, Smart People, Smart Living, Smart Mobility, and Smart Energy. DEPA's goal in the Phuket 2020 Smart City project is to develop a big data platform that aligns different elements of Smart concepts in order to develop Phuket as a smart city in terms of Smart Economy and Smart Living Community. The cross-data platform is meant to be hosted at DEPA but owned by Phuket province. Almost the entirety of Phuket's GDP (96.5%) comes from the tourism industry, making the island extremely reliant on tourism. Therefore, one of the main goals of the Smart City project is to strengthen the tourism industry while also developing a new source of income–by turning Phuket into a regional hub of digital industry that will attract investors and tourists.

The Phuket Smart City initiative aims to improve the following: 1) safety of all citizens, 2) risk management and disaster readiness, and 3) market insight and forecasting to boost tourism.

At the beginning of the project, DEPA had already identified various data sources and was preparing for the next phase of data integration into a big data platform, the Phuket Smart City Data Platform (PCDP). We were involved in the project as consultants and software architects. Our team's mission is to develop a framework and a high-level action plan on the integration of these data sources. This paper reports the proposed PCDP framework and recommendations.

II. Related Work

Previous studies [1], [2] have highlighted the importance of data platforms as the core element to develop smart cities. Existing works [3]–[5] have shown that data acquisition from multiple systems provides vast amounts of data, therefore illustrating the importance of having a centralized data platform.

With vast amounts of raw data, data mining studies show some of the difficulties and processes required when attempting to filter and collect multiple sources of data [6], [7]. Ang et al. [8] reported that with data collection and preprocessing, overall data quality is improved when compared to the raw data, therefore presenting another benefit for data to be centralized.

By utilizing big data platforms, such as Hadoop [9], it showed that not only could it collected data, but also optimize the performance and fault tolerance of distributed systems to bootstrap the PCDP [10].

Based on survey results, Xie et al. [11] have shown the need for data platforms, such as the proposed PCDP, in order to centralize and consolidate data. A modular approach for data input from various IoT (internet of things) devices is the fundamental requirement for PCDP. Inter-operability between systems provides simplicity and scalability, allowing systems to gather data from more sources, thereby enhancing the datasets and overall value of the platform [12].

III. Research Methodology

A qualitative investigation was conducted with stakeholders including leaders within DEPA and representatives from

978-1-7281-6486-1/20/$31.00 ©2020 IEEE

different sectors – hotel and spa industry, private companies, various government agencies, the police and marine agency, and public universities. The investigation included interviews, visits to facilities, and an extensive design thinking workshop. Quantitative measurements were compiled using secondary data supplied by the stakeholders. Gathered data were complemented by desk research using various sources including social media, blog posts, vendor white papers, and scientific papers.

The stakeholders recruited for our study can be grouped as follows:

1) **High impact, interested people:** must be fully engaged with and satisfied.
2) **High impact, less interested people:** must be kept satisfied and interested.
3) **Low impact, interested people:** must be adequately informed to ensure that no major issues arise.
4) **Low impact, less interested people:** must be monitored, but without excessive communication. again, monitor these people, but do not bore them with excessive communication.

IV. Results

This section describes our findings based on interviews and discussions we had with the various stakeholders. We highlight the findings in Table I. These findings were used as input in the development of the framework.

There are currently many data sources used by DEPA. Few of these data sources are interconnected, which hampers identification of interrelations between the entities. There is already a cloud hosted by Communications Authority of Thailand (CAT) [14] Telecom from which some of the data sources are fed.

Around Phuket, there are currently 14 IoT-sensors installed. These measurements depend on their location, temperature, rainfall, humidity, water level, carbon monoxide proportion, and oxygen saturation of water. The data is transferred via mobile data, is owned by DEPA, and is hosted in the CAT cloud.

The Disaster Command Center (DCC) of Phuket currently captures footages over 300 cameras, with a third of them having their footage stored for 30 days. There are two concurrent Closed-circuit television (CCTV) systems: one from the police and one owned by Phuket province. These cameras are placed at Phuket bridge and on various junctions across Phuket province. Currently, there are also plans to install cameras along the shores. The cameras on Phuket bridge have optical character recognition (a.k.a. OCR) capability to extract license plate numbers from the footage and a facial recognition option. Facial recognition results are matched on-the-fly to a given dataset but not stored. In addition to individual license plates, the system also captures aggregates, e.g., the number of vehicles passing Phuket bridge within a certain period of time (month or day). The data is stored locally at the DCC. For the PCDP, although the aggregates could be used, individual personal information cannot be utilized for privacy reasons. However, it might be possible to use individual information with anonymized number plates (e.g., by using a hashed value instead of the actual number plate). Incidents are also tracked but not in the context of the location. This might be an option to consider in the future, for example, to identify junctions with the highest number of traffic incidents. In the near future, more cameras, including those privately owned, will be integrated into the system.

DEPA also hosts a smart growth web application, currently accessible to tour operators and hotel owners to upload guests' demographic information. The website is also used to submit product and service promotions. This information is stored in the CAT cloud owned by DEPA. The promotion events are then submitted to various beacons around Phuket and can be viewed on the Smarter City mobile application.

Another potential data source lies within the Marine department of Phuket. The Vissim Vessel Traffic Management System allows for real-time capture and tracking of the location, direction, and speed of vessels. There is also a CCTV system that stores footage for two years. The buoys and sensors provide information on water level, oil spill, and land topography. All data are stored locally and currently not integrated into the data sources used by DEPA.

There are currently two new data source projects under way. One, in cooperation with visa card operators in Phuket, collects demographic information of visa card users along with the volume of transaction details. The other project aims to source consumer demographics and usage details of the free public Wi-Fi service provided by CAT Telecom. It might also provide information about the location of the user, a unique identifier per device (MAC-address), the device type, and the person registering.

Two more data sources planned for integration are the Phuket Provincial Statistical Office Open Data and energy consumption statistics from the Provincial Electricity Authority in Phuket.

As our analyses and work on business value propositions have shown, it will be necessary for the PCDP to also tap into commonly available data sources such as news platforms, statistics, social media, tourism industry portals, and other public sources. This does not mean that the current existing data sources are not valuable. In fact, the Phuket-specific data sources are the differentiator between the PCDP and other similar big data platforms. However, in order to provide real business value, it will be necessary to add other data sources to form a broader picture of the available datasets.

A. *PCDP Architecture*

The proposed PCDP serves the purpose of integration of several data sources, transformation, consolidation, cleansing, and storage of data, up to an efficient way of provision for analysis and interpretation. While it would be possible in some cases to perform analytics in a single source system, this operation might create additional load on it. This additional stress might have a negative impact on the operation of the source system. In addition to data being available in different formats

TABLE I
SUMMARY OF FINDINGS FROM INTERVIEWS

Key Stakeholders	Tourism Industry	Private Organization	Public Organization
Pains / Needs	**Phuket Tourism Authority** Hotel Industry - Inaccuracy of sales (occupancy) records or operation (workers) forecasting - Lack of customer information for: 1) executing targeted marketing campaigns, and 2) delivering customer centered services (e.g., type of food) - Manual market rate monitoring - Inability to do benchmarking - Lack of information for investment planning - Manual social media web crawling (i.e., lack of sufficient economic means to get information from search engines, travel websites, etc.) **Tour Agencies** - Illegal tour operations (inefficient government policy) - Lack of multilingual tour guides especially Chinese, Russian, and Korean - Lack of graduates specializing in tourism industry - Lack of analytics tools to process data surveys. (Agencies find it hard and time consuming.) - Inability to determine prime languages when creating tour guides due to lack of information on tourist statistics	**Phuket City Development [13]** - Varied information from source to source - Disjointed information complicates decision-making - High risks on investment due to lack of market information - Lack of sufficient information on availability of land in Phuket - Inability to integrate variables like flooding to make smarter investment decisions **Software Entrepreneur** - Lack of a strong developer community compared to Bangkok - Scarcity of specialized IT engineers - Lack of market analysis to see where they could grow their services (local and national) - Lack of comprehensive customer data to generate new businesses - No access to real-time data	**Marine Department** - Inability to track private or international boats - Inability to integrate other valuable sources of data into their system - Inability to identify owner or passengers in a boat **Researchers** - Limited access to data - Inaccuracy of data at the Disaster Command Center **Disaster Command Centre (DCC)** - Lack of system integration between police and ministry databases - Basic analytics done

from different systems, some source systems may only store historical data to a limited extent. Fig 1 shows the architecture of PCDP, including the following layered structures (each layer will be described in the following sections):

- Ingestion/Staging: After importation or extraction from the source system, the data is kept in an unchanged format.
- Curation/Warehousing: At this stage, data is transformed, cleansed, and often converted into an application-neutral format–the foundation for integration with data from other data sources.
- Analytics/Data Marts: Depending on its intended usage, data at this layer can be either granular or highly aggregated. It conforms to the output of the interpretation.
- Provision/Presentation: This layer is responsible for the provision of data in various formats to consuming applications, or to data representation and presentation towards the end user.

B. PCDP Ingestion Framework

Data ingestion is the process of obtaining and importing data for immediate usage or storage in a database. To ingest is to take in or absorb something. Data can be streamed in real time or ingested in batches. When data is ingested in real time, each data item is imported as it is emitted by the source. When data is ingested in batches, data items are imported as discrete chunks in periodic intervals of time. An effective data ingestion process begins by prioritizing data sources, validating individual files and routing data items to the correct destination.

Fig. 2 shows the elements in PCDP ingestion framework, including:

- **Data Asset:** data in any form–as a single file, document stream, elastic search index, or an entire database
- **Data Asset Owner, a.k.a. the Publisher:** person or organization who has ownership of a data asset to be published in the PCDP
- **Drop Zone (DZ):** staging part of the PCDP where data is initially deposited or “dropped” by data asset owners
- **Landing Zone (LZ):** area of the PCDP where data can be accessed by authorized users
- **Hadoop:** formally referred to as Apache Hadoop, is an Apache software foundation (ASF) project and open source software platform for scalable, distributed computing. Hadoop can provide fast and reliable analysis of both structured and unstructured data. Given its capabilities to handle large datasets, it is often associated with the term ‘big data’. We refer to Hadoop here in its most general meaning. Hadoop is an ecosystem of software components that use the Hadoop Distributed File System (HDFS). In this sense, Spark [15] is a part of Hadoop.
- **HDFS:** a sub-project of the Apache Hadoop project. This ASF project provides a fault-tolerant file system that is designed to run on commodity hardware.
- **HDFS DZ:** the area of HDFS where the user can store any file. The HDFS DZ is accessible from a web browser

Fig. 1. PCDP Architecture Overview

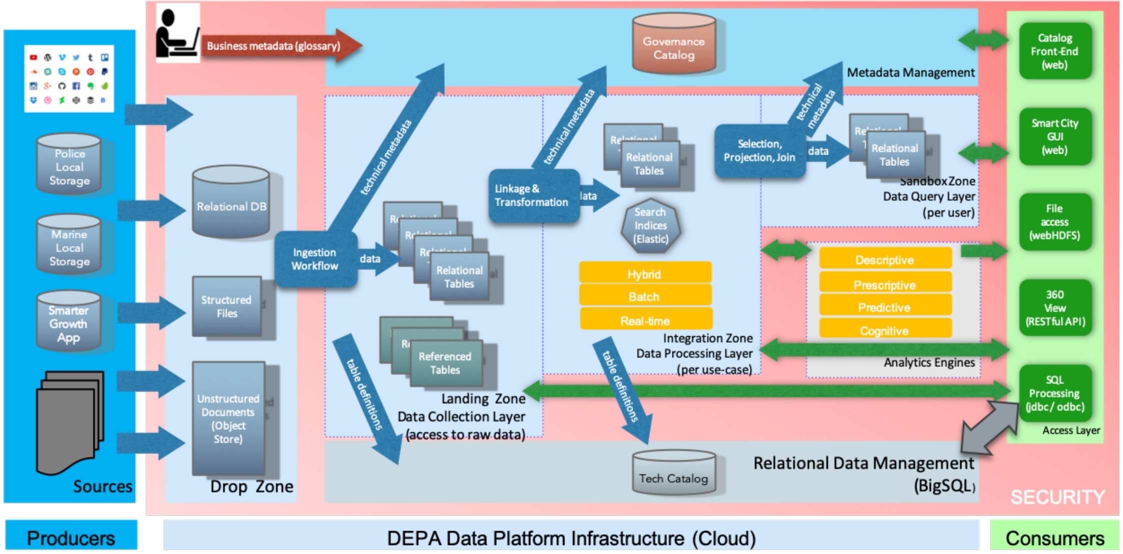
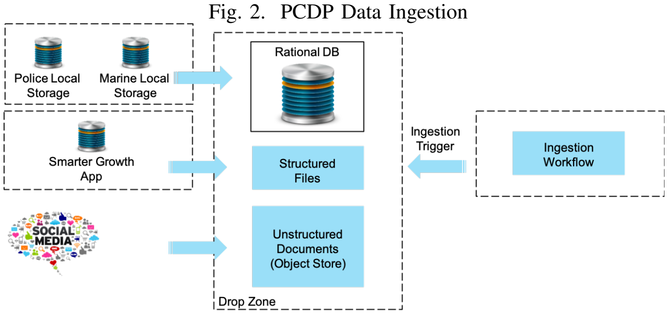


Fig. 2. PCDP Data Ingestion



or programmatically via the WebHDFS protocol using a suitable HTTP client (e.g., curl).

- **Apache Kafka:** open-source publish-subscribe messaging system designed to provide quick, horizontally scalable, and fault-tolerant handling of real-time data feeds. Unlike traditional enterprise messaging software, Kafka is able to handle all of the data flowing through a company in near real-time.
- **Elasticsearch:** a search engine based on Apache Lucene. It provides a distributed, multitenant-capable full-text search engine with a HTTP web interface and schema-free JSON documents. Elasticsearch is developed in Java and is released as open source under the terms of the Apache license. Official clients are available in Java, .NET (C#), Python, Groovy, and many other languages. Elasticsearch is the most popular enterprise search engine followed by Apache Solr, which is also based on Lucene.
- **Catalog:** a metadata repository, where data is registered after being moved to the LZ. Users of the PCDP use the catalog to browse and understand the available data assets.
- **Consumer:** any user who is authorized to access or consume a data asset in the LZ
- **Groups:** Access to the DZ and LZ are determined by membership in asset-specific groups.

Designing a successful data ingestion framework is a critical task, requiring a comprehensive understanding of the technical requirements and business decisions to fully customize and integrate the framework for enterprise-specific needs. Data ingestion is composed of the steps described below.

Data asset owners make initial clarifications with the governance expert within the PCDP governance organization. Once the asset owner and PCDP governance have agreed to onboard the data asset, the PCDP provides administrative rights to those performing the data transfer into the DZ. The exact nature of these rights differs depending on ingestion type but typically include a 'producer right' to access the DZ and trigger data ingestion. The right to read data from PCDP is managed by a 'consumer right' that is distinct and separate from the 'producer right'. The right to ingest data does not automatically include the right to retrieve data from PCDP. If the data asset is relational, then the PCDP governance team initiates the creation of a DZ database. The data asset producer is notified of the creation of the DZ database and is provided with the necessary configuration details to it. In order to upload files, the data asset owner must have access to the HDFS DZ. Data is ingested by the person(s) in the producer group. The various ways by which the ingestion process is triggered will be described later. An ingestion trigger starts a workflow which copies the data from the DZ into the LZ in the PCDP, updating the catalog with technical metadata about the data asset. The workflow creates policies which authorize the consumer group for BigSQL artifacts associated with data in the LZ, as well as read access to the underlying file system resources depending on the ingestion type.

An ingestion trigger is used to start the copying of data from the DZ into the LZ of the PCDP. The ingestion process can work on a single database table in the case of relational data, on a set of files located in a directory in the case of file data, and on a document type in an elastic search index. It carries out the following tasks:

- Copying of data from the DZ to the LZ.
- Making relational and structured data from Elasticsearch available on BigSQL. Access permission is granted to PCDP developers and to all users in the consumer group of the corresponding data source.
- Making raw data files available.
- Importation of technical data into the governance catalog (Optional).

The ingestion of relational data is outlined in the following process:

1) For every relational data source, a dedicated database is created in the PCDP DZ database system. Access rights to this database is given to the 'producer' group. Membership in the producer groups is maintained by the PCDP governance team.
2) The producer loads data into the corresponding DZ database by using the data stage, shell scripts, or any other tool.
3) Once the data is available in the DZ database, the producer triggers the ingestion workflow which copies a single database table into the LZ and, optionally, updates technical metadata in the governance catalog.

Data becomes available in BigSQL after the ingestion process has been completed. If the table is being updated, only old data will be available during the ingestion process until the process is completed, at which point the new data also becomes available. When the table structure is changed, the table is immediately invalidated and the old data is no longer available during the ingestion process.

C. Data Curation

Data curation is the process of turning independently created data sources (structured and semi-structured data) into unified datasets ready for analytics by using domain experts to guide the process. It involves the following:

- Identification of data sources of interest (whether from inside or outside of the enterprise)
- Data verification (to ascertain its composition)
- Cleaning the incoming data (e.g., 99999 is not a legal zip code)
- Data transformation (e.g., from European to US date format)
- Integration with other data sources of interest
- Deduplication of the resulting composite dataset

The objective of data curation is to select the most relevant information for consumers, and to classify, enrich, and distribute it in ways that can be readily consumed. The curation process uses a number of curators in the content classification process, working over a number of data sources. A curator is an analyst who collects, aggregates, classifies, normalizes, and analyzes raw information coming from different data sources. Because of the high volume and near real-time influx of analyzed information, data curation is a big challenge inside the organizations and the use of automated tools plays an important role in this process.

D. Data Analytics

The PCDP framework supports four analytics types as follows:

- Descriptive - used to understand at an aggregate level what is going on with a particular business, based on historical data
- Predictive - used to identify the likelihood of future outcomes
- Prescriptive - used to provide users with advice on what action to take
- Cognitive - used when there are massive amounts of information (structured and unstructured) and multiple data sources involved

E. Presentation and Data Delivery

The main component of the presentation is a dashboard, which graphically represents analysis results in a condensed and comprehensible format. Various metrics and key performance indicators are used to enable viewer insight into complex dependencies.

The dashboarding process involves the identification of relevant information, definition of the audience, and the presentation. As these steps are highly interdependent (e.g., selection of audience influences the relevance of information), they are most likely iterative.

Users are only willing to include dashboards into their decision process if they trust the information they receive. One way to establish confidence is to give users the option to alter certain parameters so users can develop an intuitive feeling (What-If Analysis). Graphical presentation can help in identifying facilitate the identification of data source problems or flawed assumptions in modeling that might be difficult to detect in tabular format [16].

Data presentation is only valuable if the decision maker can act upon it. Especially in the case of big data, poor visualization often leads to the 'beautiful hairball problem' [17] : complex and impressive graphs, but incomprehensible, with little filtering and annotation.

Interpreting results also involves having insights into the assumptions made and being able to drill down in order to retrace and verify them. Other than just being confronted with graphs and figures, users should be able to understand why they are seeing the current results.

Another component of the dashboard is the creation of alerts. Dashboards can provide an interface for users to be able to:

- specify thresholds
- define threshold level calculation, e.g., value-based vs. percentage-based alerts
- trigger threshold-based events

Other types of information representation include:

- map components (geographical data), e.g., tourists per municipality
- calendars

V. Conclusion and Future Work

This study covers recommendations regarding the technical composition of the PCDP. This study is accompanied by a slide deck, which further elaborates and visualizes one specific business value proposition for each of the key areas. Numerous interviews, design thinking workshops, and discussions supported the identification of the data landscape as well as the potential application areas and business values for the PCDP. The major recommendations for the PCDP framework can be split into a functional and an administrative section. The functional section addresses technical core functionalities, such as:

- Data Ingestion: how data is imported into the platform
- Data Curation: how data from separate sources are integrated into a unified dataset
- Analytics: how insight can be collected and generated from a data pool
- Presentation and Provision: how data and results can be shared with users

The administrative section covers the fields of:

- Governance: enabling and enforcing desired behavior and limiting unwanted actions
- Security and Privacy: controlling access to data and protecting sensitive information
- Data Quality Management: data cleansing and consolidation
- Organization Structure: description of necessary roles and responsibilities involved in running a big data platform

One of the key findings is that in order to provide real business value to its stakeholders, the PCDP needs to be enhanced with additional publicly available data sources. For many use cases, including tourism, digital industry, and safety, social media information plays an important role. Therefore, the PCDP should be enhanced with social media feeds and analytics capabilities. This does not mean that the Phuket-specific data sources are worthless–they are the main differentiator between the PCDP and other similar initiatives. Other than that, broader data sources such as newsfeed extraction, open statistics, and social media must act as connecting elements between existing data sources to make the PCDP a valuable source of information.

References

[1] A. Ojo, E. Curry, and F. A. Zeleti, “A tale of open data innovations in five smart cities,” in *2015 48th Hawaii International Conference on System Sciences*, 2015, pp. 2326–2335.

[2] H. M. Nguyen, J. Byun, K. Kwon, J. Han, W. Yoon, N. Lee, H. Kim, N. Pham, D. Kim *et al.*, “Oliot-opencity: Open standard interoperable smart city platform,” in *2018 IEEE International Smart Cities Conference (ISC2)*, 2018, pp. 1–8.

[3] S. Kazi, M. Bagasrawala, F. Shaikh, and A. Sayyed, “Smart e-ticketing system for public transport bus,” in *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, 2018, pp. 1–7.

[4] A. Souza, M. Figueredo, N. Cacho, D. Araújo, J. Coelho, and C. A. Prolo, “Social smart city: A platform to analyze social streams in smart city initiatives,” in *2016 IEEE International Smart Cities Conference (ISC2)*, 2016, pp. 1–6.

[5] S. Xiong and B. Ye, “Analysis on the development path of smart city in the era of big data,” in *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, 2019, pp. 177–181.

[6] S. K. Dwivedi and B. Rawat, “A review paper on data preprocessing: A critical phase in web usage mining process,” in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, 2015, pp. 506–510.

[7] S. Sharma and A. Bhagat, “Data preprocessing algorithm for web structure mining,” in *2016 Fifth International Conference on Eco-friendly Computing and Communication Systems (ICECCS)*, 2016, pp. 94–98.

[8] Q. Ang, Z. Liu, W. Wang, and K. Li, “Explored research on data preprocessing and mining technology for clinical data applications,” in *2010 2nd IEEE International Conference on Information Management and Engineering*, 2010, pp. 327–330.

[9] “Apache hadoop,” https://hadoop.apache.org/, (Accessed on 03/27/2020).

[10] E. Sivaraman and R. Manickachezian, “High performance and fault tolerant distributed file system for big data storage and processing using hadoop,” in *2014 International Conference on Intelligent Computing Applications*, 2014, pp. 32–36.

[11] Y. Xie, J. Gupta, Y. Li, and S. Shekhar, “Transforming smart cities with spatial computing,” in *2018 IEEE International Smart Cities Conference (ISC2)*, 2018, pp. 1–9.

[12] N. Villanueva-Rosales, L. Garnica-Chavira, V. M. Larios, L. Gómez, and E. Aceves, “Semantic-enhanced living labs for better interoperability of smart cities solutions,” in *2016 IEEE International Smart Cities Conference (ISC2)*, 2016, pp. 1–2.

[13] “Home page - ,” http://www.pkcd.co.th/, (Accessed on 03/27/2020).

[14] “Cat telecom,” https://www.cattelecom.com/coverpage/start.php, (Accessed on 03/27/2020).

[15] “Apache spark - unified analytics engine for big data,” https://spark.apache.org/, (Accessed on 03/27/2020).

[16] J. G. Stadler, K. Donlon, J. D. Siewert, T. Franken, and N. E. Lewis, “Improving the efficiency and ease of healthcare analysis through use of data visualization dashboards,” *Big Data*, vol. 4, no. 2, pp. 129–135, 2016.

[17] “Are you tangled in a big data hairball?” https://www.forbes.com/sites/lisaarthur/2013/08/01/are-you-tangled-in-a-big-data-hairball/7b8705bc7776, (Accessed on 03/27/2020).