



INTELLIGENT TRANSPORT SYSTEM
ASSOCIATION OF MALAYSIA

ITS SEMINAR AND EXHIBITION 2017
**DRIVING ITS
TO A NEW NORMAL**

Connexion@Nexus
Bangsar South, Kuala Lumpur, Malaysia
21 - 23 February 2017

Organised by



In Collaboration



Supported by





INTELLIGENT TRANSPORT SYSTEM
ASSOCIATION OF MALAYSIA

Analytics with Big Data for Intelligent Transport System

Presented by:
Dr. Keeratpal Singh
Principal Data Scientist
22nd Feb 2017
Session B4,
Bangsar South, Kuala Lumpur



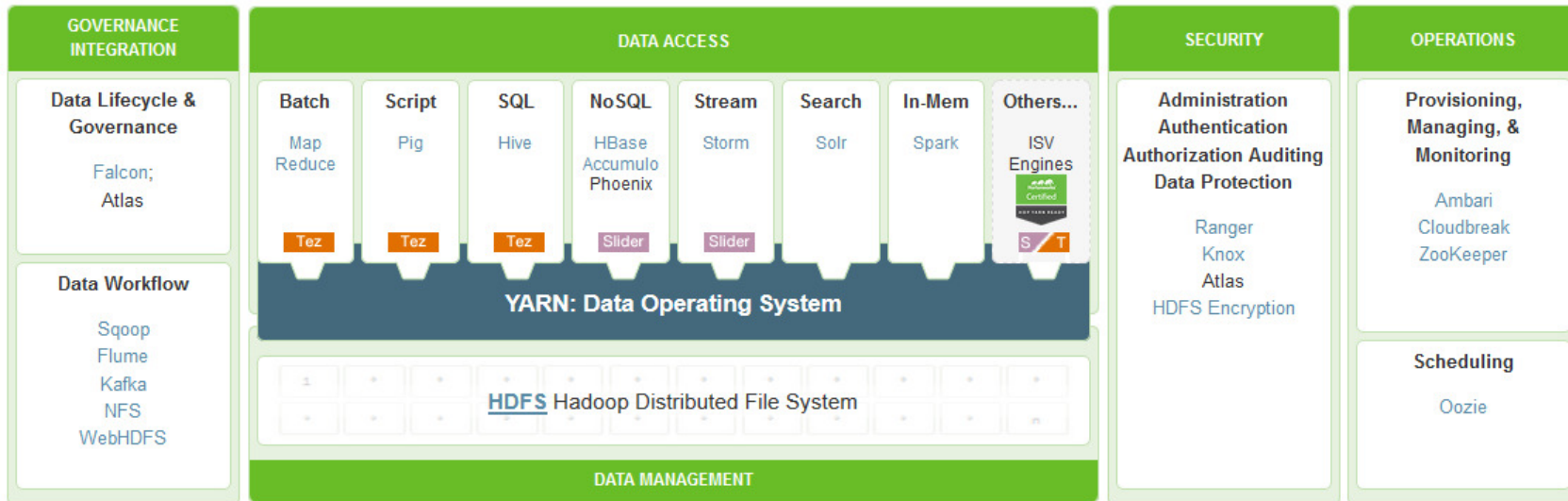
Content of Paper



- Applying analytics and visualization tools to guide decision makers in solving transportation challenges and addressing customers' pain points.
- Generating real or near real-time analysis with variety of data to drive transportation system's directions
- Incorporating simulation and innovative techniques for predictive and prescriptive analytics
- Bridging Mobile Applications as IoT with Big data processing for real time analytics in transportation system
- Data Cluster: Extracting customer information, validating datasets, analyzing and mining, discover pattern and trends, visualization and creating value for the customer, and finally protecting data.
- The process of visualization of time series forecast for large datasets in transportation sector with billion records
- Comparison of processing time for types of search query with our home grown analytic tool using RDBMS, Big Data Hadoop nodes and GPGPU DB



Big Data Stack: Example



Reference and Source: Hortonworks Data Platform



Big Data Platform: Example



Analytic Applications

BI /
Reporting

Exploration /
Visualization

Functional
App

Industry
App

Predictive
Analytics

Content
Analytics

IBM Big Data Platform

Visualization
& Discovery

Application
Development

Systems
Management

Accelerators

Hadoop
System

Stream
Computing

Data
Warehouse

Information Integration & Governance

Reference and Source: IBM

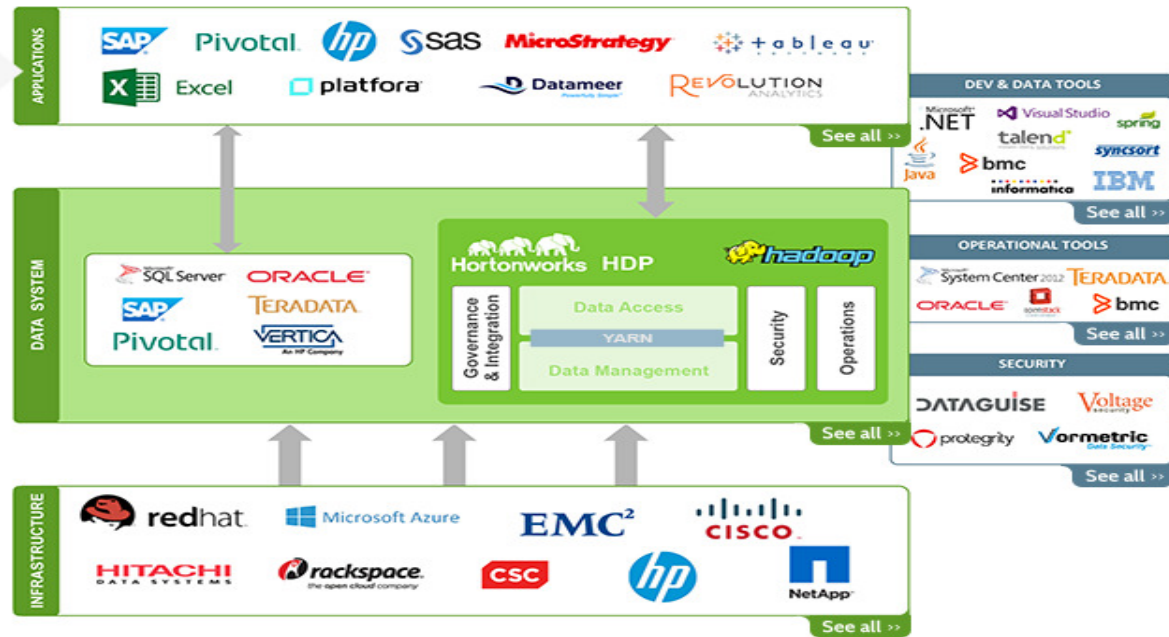


Business Intelligence and Tools for Analysis



A robust ecosystem of technologies and tools augments HDP for enterprise usage

Business Intelligence and Analytics
 All the major BI vendors offer Hadoop integration, and specialized analytics vendors offer niche solutions for specific data types and use cases.



Data Management and Tools
 There are many partners offering vertical and horizontal data management solutions along side Hadoop, and there are numerous tool sets – from SDKs to full IDE experiences – for developing Hadoop solutions

Infrastructure
 While Hadoop is designed for commodity hardware, it can also run in appliances or other storage solutions, on-premise and in the cloud, or in a virtualized environments.

Reference and Source: Hortonworks



Variety of Data



- Unstructured Data
- Government Data
- Private Cloud
- Enterprise Cloud
- Spatial and GIS Data

Extracting the value of Data



E
x
t
r
a
c
t

D
a
t
a

I
n
t
e
r
p
r
e
t

D
a
t
a

V
a
l
u
e

Data Harvester

Data Cleanser

Data Harmonizer

Data Anonymizer

Data Modeler

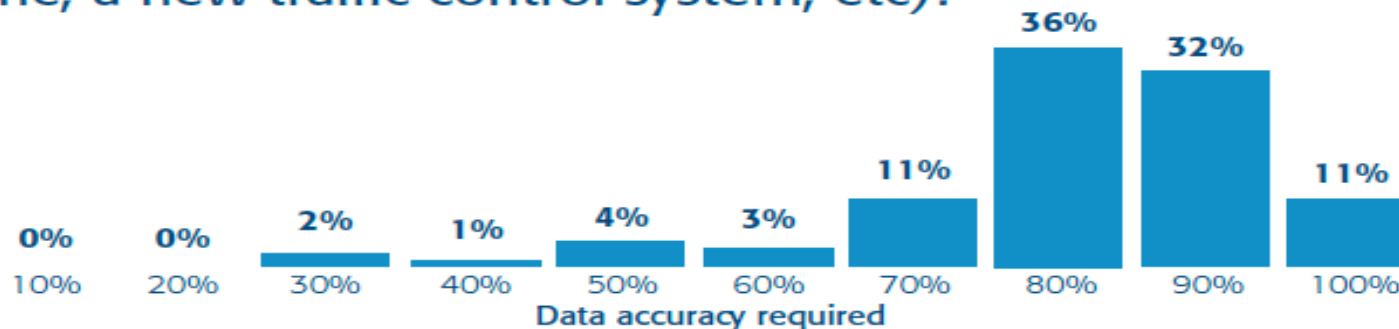
Data Consumer Channels

1. Extracting customer information
2. Cleansing and transformation
3. Validating datasets and harmonizing terminology
4. Protecting data
5. Analyzing and mining to discover pattern/trends to create models and data warehouse
6. Visualization through dashboards or mobile channels to create value for the customer



Integrity of Data

FIG. 7 In order to make a business decision you need to feel confident in the data; however it may not be worth the effort to ensure data is 100 per cent accurate. At what percentage of data accuracy would you be comfortable in terms of using it to make an important decision (e.g. a major marketing campaign, the launch of a new product line, a new traffic control system, etc)?



One-hundred per cent data accuracy is a goal that is out of reach for a majority of organisations given budget limitations. However, the fact that a majority of respondents told us that somewhere around 80 to 90 per cent was “good enough” was telling.

Reference and Source: Computing Research, Big Data Review 2015



Analysis of data: What are the events related to users' location



- Mobile Phone as collector uses GPS and network (IoT)
- Backend using Big Data Engine with Apache Solr for Fast Search (Big Data)



Mobile as IOT

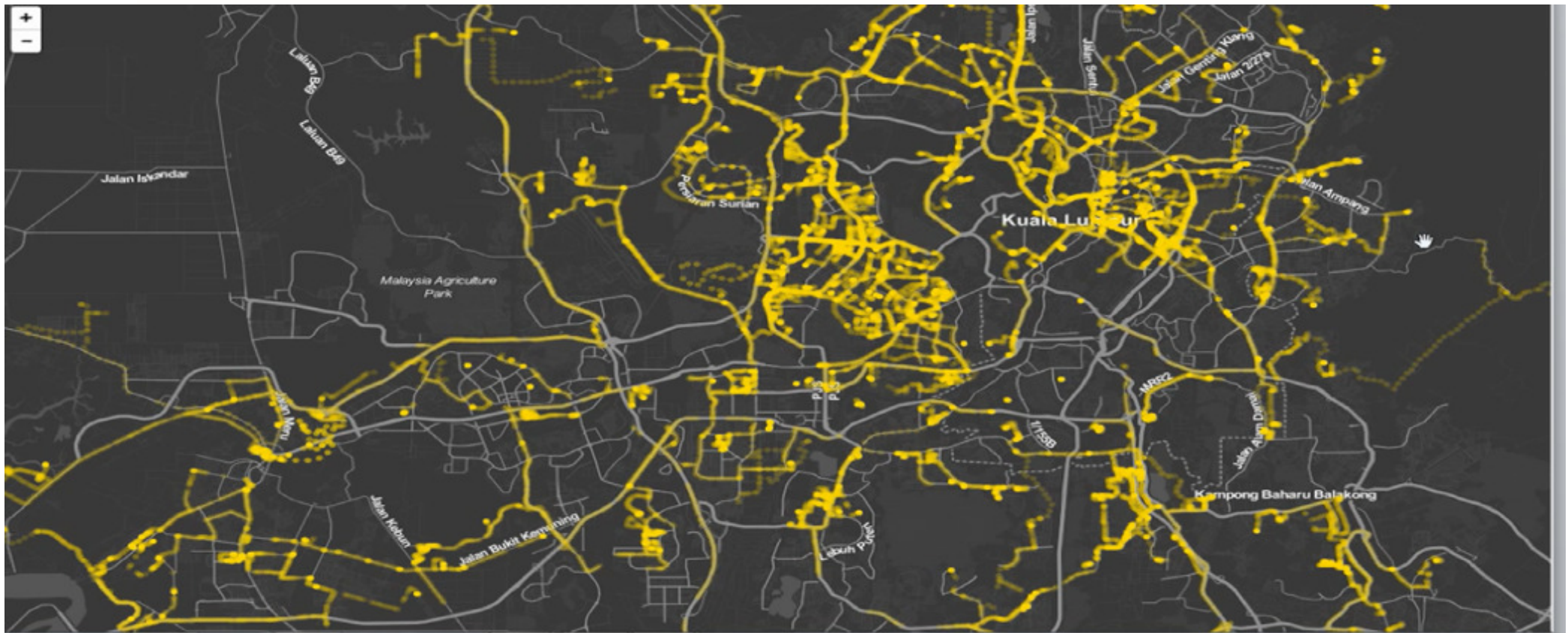


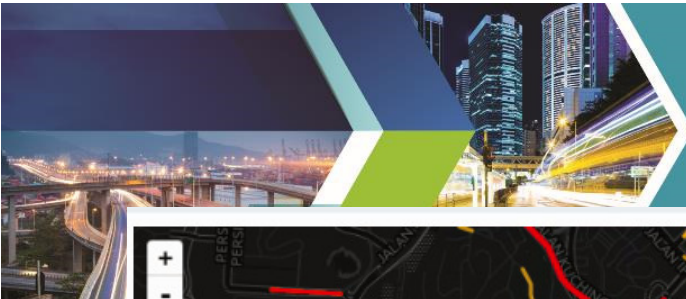
Big Data Reporting Engine

Using Apache Solr: Fast Search



Fleet Data (Demo Spatial) million of points on a map and analyzing each point in real time





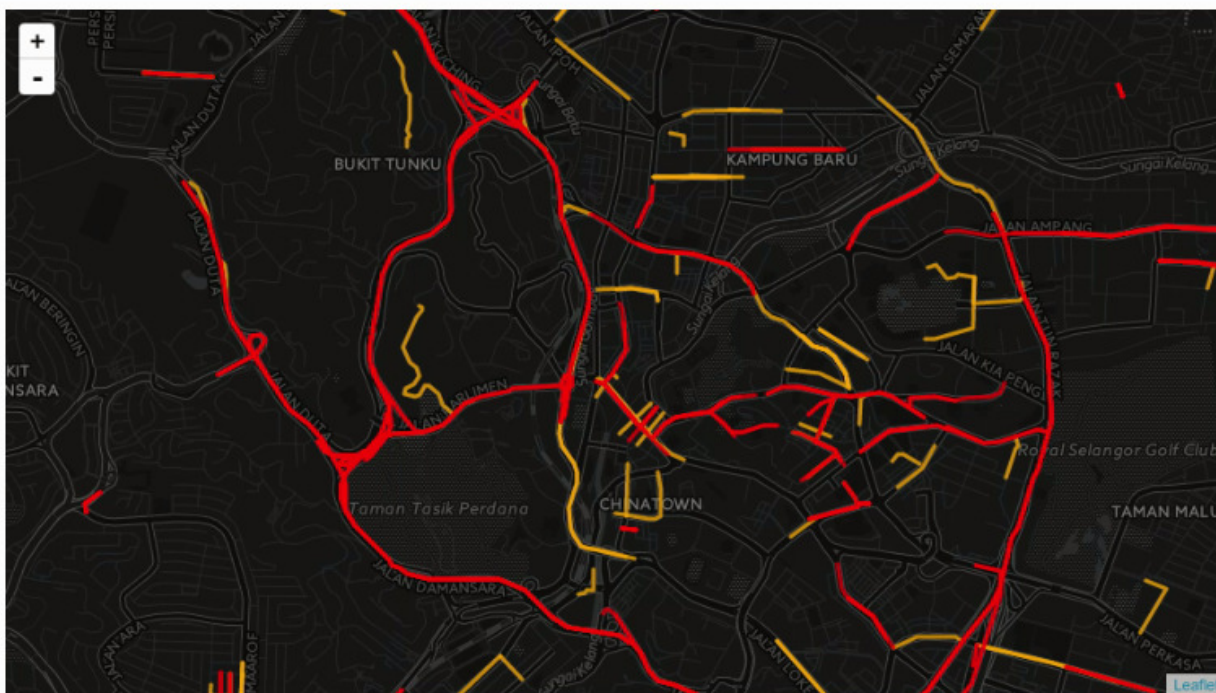
Waze Data: Traffic Jam by city



Waze users' concentration per city in March, 2016.

This analysis illustrates traffic pattern per hour.

It shows less traffic around early noon.

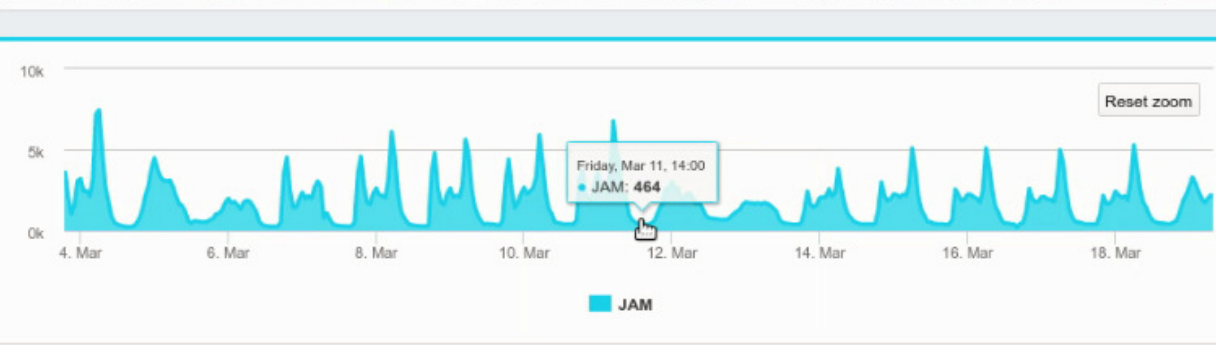


Most Traffic Jams

	city	total
1	Kuala Lumpur	4416564
2	Petaling Jaya	1346393
3	Puchong	980239
4	Klang	669715
5	Shah Alam	656053
6	Subang Jaya	335915
7	Kajang	323340
8	Seri Kembangan	184064
9	Sungai Buloh	181530
10	Batu Caves	180430

Select City

Kuala Lumpur

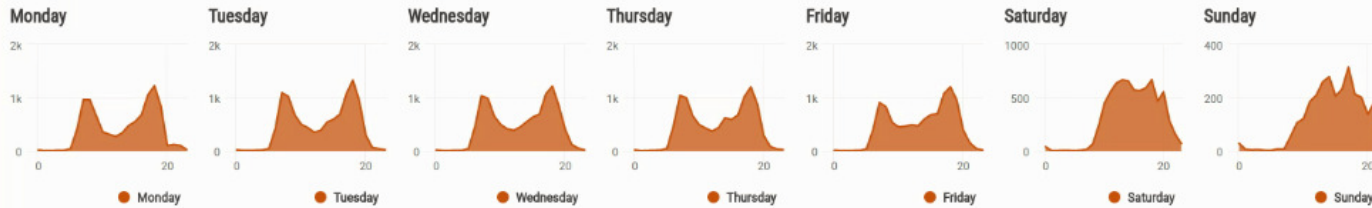




Waze Data: compare traffic for different streets



"E11 LDP"



Select Region:

Petaling Jaya

Select Street under Petaling Jaya :

"E11 LDP"

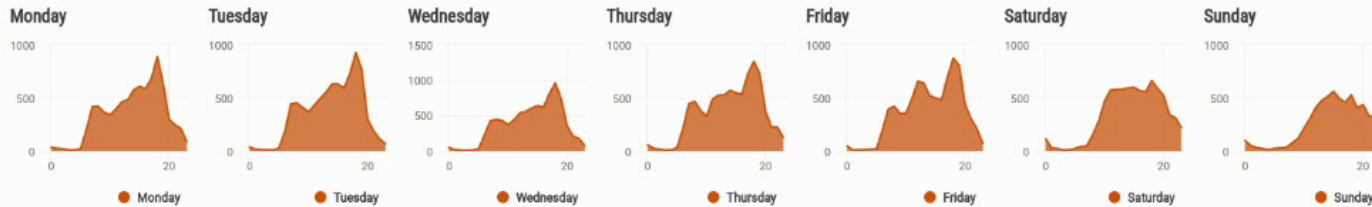
Select Street under Petaling Jaya :

"Persiaran Surian"

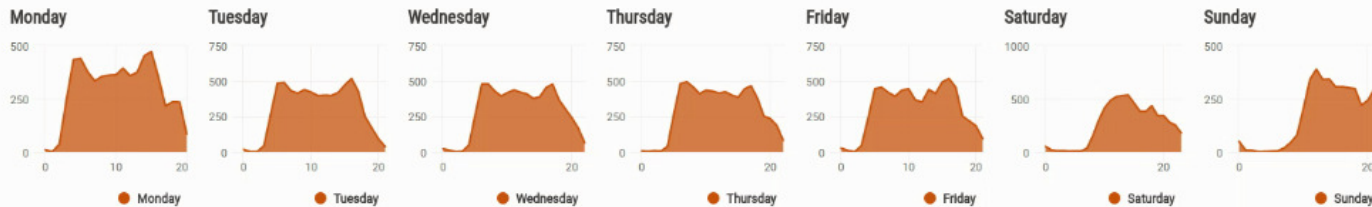
Select Street under Petaling Jaya :

"2 Lebuhraya Persekutuan"

"Persiaran Surian"



"2 Lebuhraya Persekutuan"



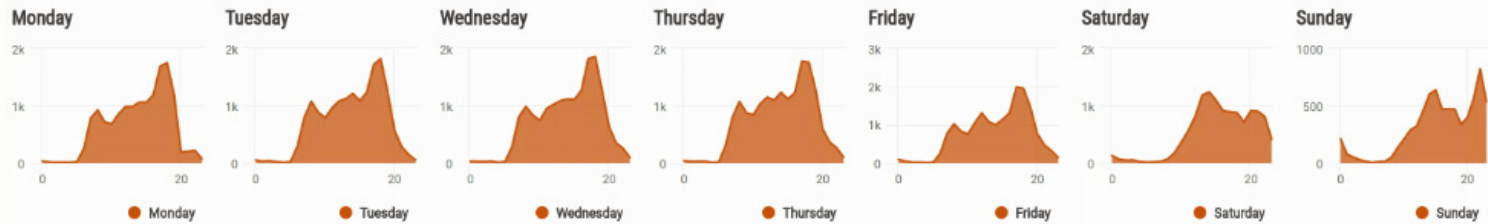


Waze Data: compare traffic for different streets in KL

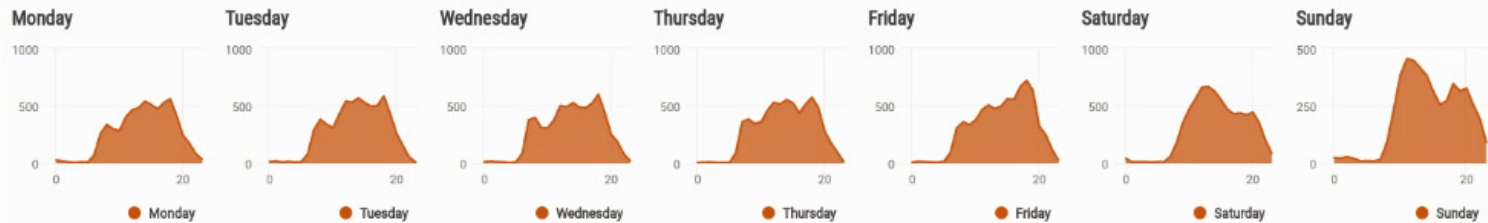


The importance of mobile users' using applications or installing IOT sensors for traffic planning in KL city

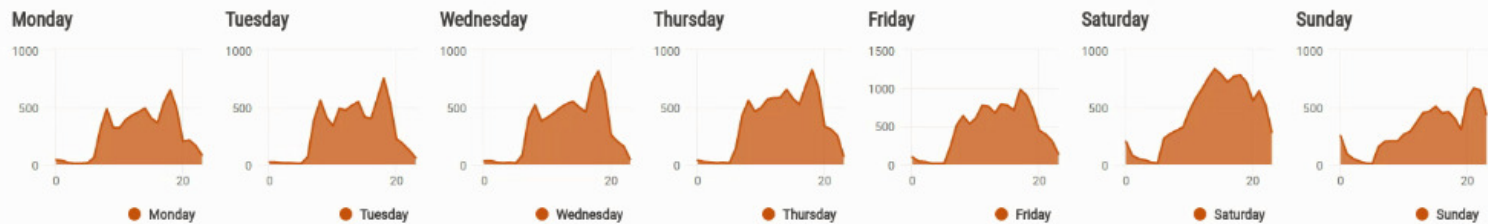
"Jalan Tun Razak"



"Jalan Pudu"

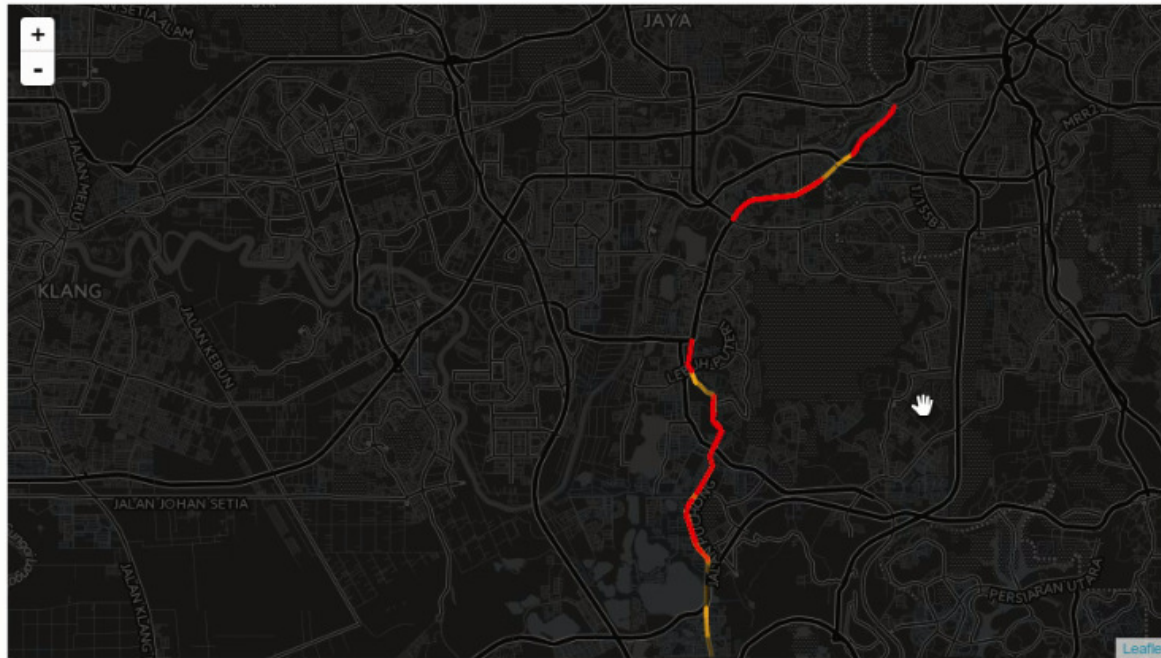


"Jalan Sultan Ismail"



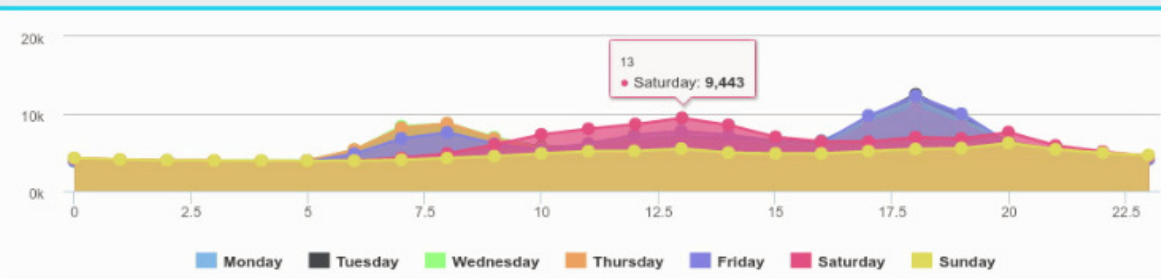


Waze Data: Hourly and weekly pattern for a street in Puchong



Select Region:
Puchong

Select Street under Puchong:
"B11 Jalan Puchong"





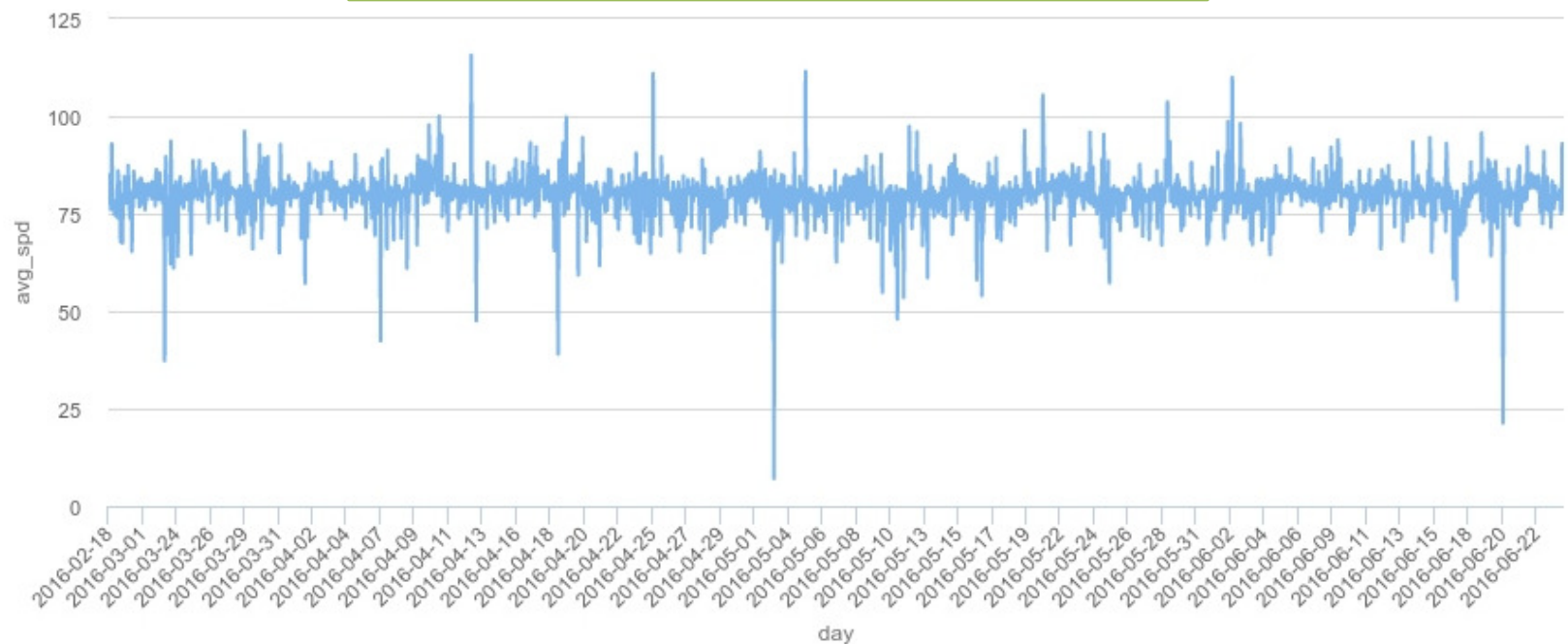
Radar sensor on an anonymous street in Malaysia



Predicted speed : 80

Next Hour : 80

Hourly averaged speed calculated from sensor



Sensors to monitor traffic patterns.

IoT usage for Road transportation planning and safety monitoring



Big Data Processing on airline dataset



- Airlines Data in the USA with about 123 million and its inflated 1.23 billion records are scooped into HDFS.
- Figure on left shows the query count and query return time in seconds.
- Figure on right shows the output after processing through time series forecasting

```
[hadoopMaster:21000] > select count(*) from airline_dataimp;
Query: select count(*) from airline_dataimp
+-----+
| count(*) |
+-----+
| 123534969 |
+-----+
Returned 1 row(s) in 1.39s
[hadoopMaster:21000] > select count(*) from airline_dataimp10x;
Query: select count(*) from airline_dataimp10x
+-----+
| count(*) |
+-----+
| 1235349690 |
+-----+
Returned 1 row(s) in 10.23s
[hadoopMaster:21000] >
```

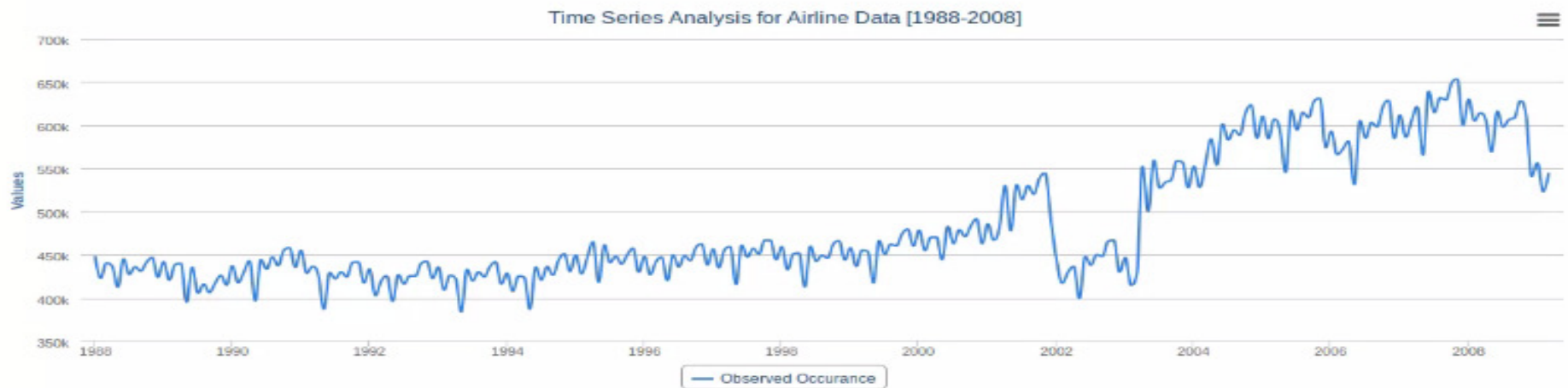
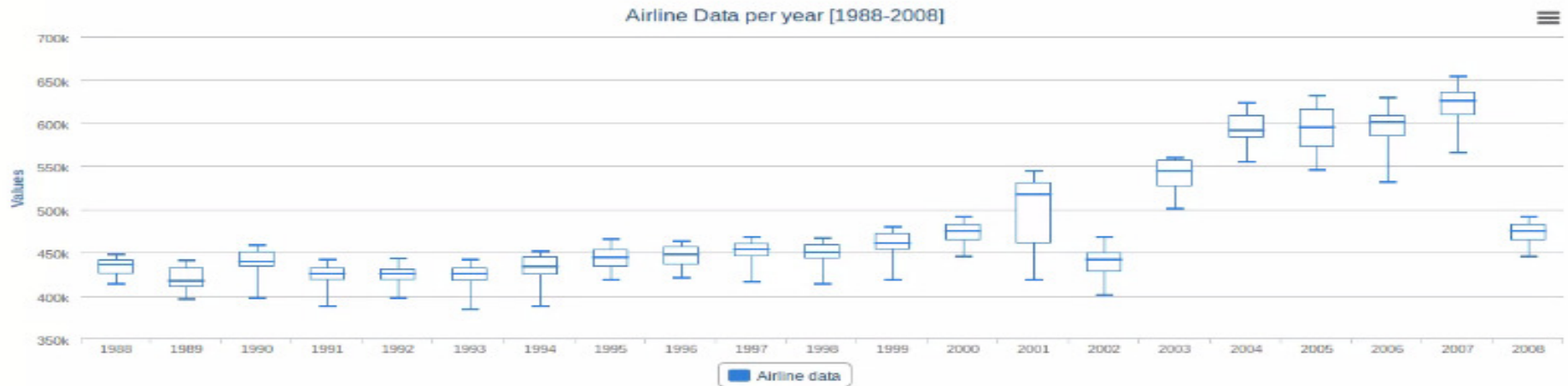




Big Data in Airlines: Airlines Data in the USA 1988-2008



with 123 million and 1.23 billion records stored on HDFS

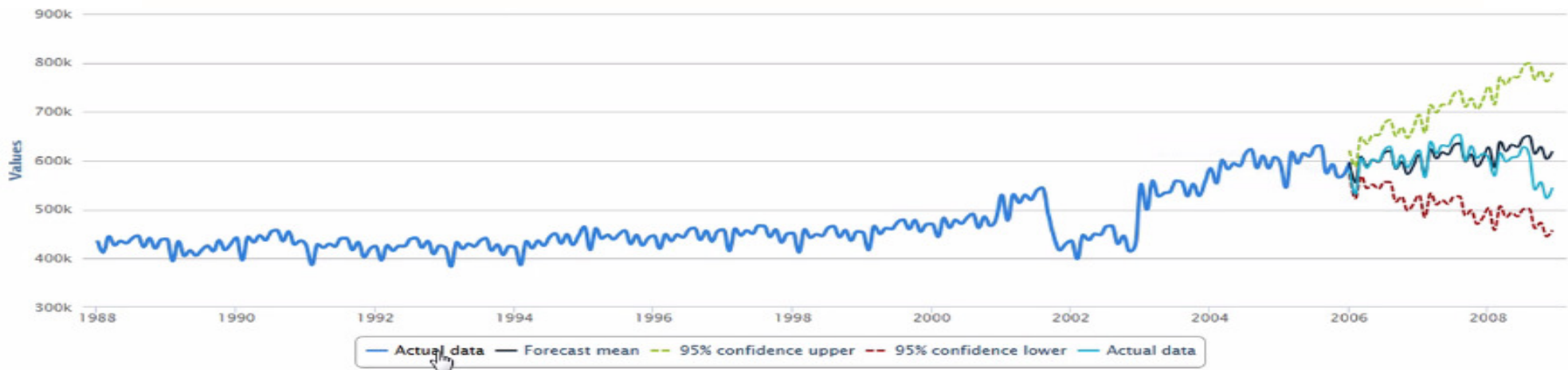




Statistical Analysis of Airline Dataset (Passengers USA)



Testing forecast of Airline Train Data (1988-2005) with Test Data (2005-2008)





Analyzing Departure Delay with Airline Data



Analyzing Departure delay with 123 million and 1.23 billion records. Which day and month has higher delays? What measures to take?





Comparison of processing time for types of search query using RDBMS, Big Data Hadoop/Impala cloud nodes and GPU DB



- 6 types of searches for the different environment setups using our middleware framework and the Mi-BIS presentation layer.
- The average time for each query type was analyzed for Postgres database with 8GB RAM server and 96GB RAM server
- Hadoop with Hive (7 nodes), Impala with Hive (7 nodes) and GPU server processing were setup to compare the processing time (in seconds).
- The result shows the performance comparison of various setups for real time Analytics Processing of Big Data in the Health sector, using Mi-BIS Presentation Dashboard to analyze ~120 millions of records in HDFS and postgres (RDBMS) servers.
- GPU parallel DB processing takes the shortest time to process ~120 million of records and the cost is also cheaper than implementing 7 nodes of Hadoop or SQL on Server or Cloud VM



Comparison of processing time for types of search query using RDBMS, Big Data Hadoop/Impala cloud nodes and GPU DB



No	Description of Search Query vs. Average Processing Time (seconds)	SQL (8GB/4Core)	SQL (96GB/48Core)	Hadoop-Hive	Impala-Hive	GPU-Parallel DB
1	Selecting sum from one column of 120 million records	1466.7s	218.7s	347.6s	3.7s	0.3s
2	Selecting a name column, counting the name and ordering by top 10 names	7901s	1612s	505s	64.2s	NA
3	Selecting state code, years from hospital patient records with one disease code selected, group by years and state code, order by years and state code	1464.7s	103.6s	383.5s	3.5s	3s
4	Selecting state, years, disease name from hospital patient records where one disease name type is selected and joining disease code with disease name and state code with state names, grouping by years, states and disease names, ordering by years and state code.	1688.7s	102.7s	N/A	2.9s	1.6s
5	Inserting the results of Selecting state code, years from hospital patient records with ALL diseases type, group by years and state code, order by years and state code	Failed	7878s	557.3s	10.1s	N/A
6	Selecting state, years, disease name from hospital patient records where three disease name types are selected and joining disease type with disease name and state code with state names, grouping by years, states and disease names, ordering by years, states and disease names	1893s	704s	N/A	3.7s	6.3s

- SQL was not able to compute on big data especially for real time analytics (218- 1466 times slower)
- Hadoop with Hive is not suitable for real time processing and would only be useful for batch processing of big data.
- Impala-Hive is as beneficial as GPU for general queries and could be used to complement in the hybrid parallel processing
- Impala Hive is faster compared to GPU Parallel DB when there are multiple tables to be joined and with huge strings operations to be performed.



Application of IOE (Internet of Everything) and IoT



The real value of the **Internet of Everything** lies in the value of connections among **people, process, data, and things**, not simply in the sheer number of things that are connected.

When your car becomes connected to the **Internet of Everything**...



...**more numerous, valuable, and relevant connections** with other cars, stop signs, your home, and even the road itself will make your driving experience safer, more fun and informed, and even more efficient.

It's the **connections** that matter most.

The Internet of
EVERYTHING

#InternetofEverything
#IoE



Reference and Source: Cisco



Concept of Interconnect of Everything with Big Data in transport system



- Machine Intelligence with **limitless communication of devices** whether on internet or non internet medium as long as network communication such as RF, radar or optical networks exists
- Ability to intelligently **automate machines' behavior** with historical data trends to optimize gains and increase automation using **machine learning** with machine intelligence **models**.
- Big data **processing** with appropriate tools (such as data mining, web scraping, using R and Python with API such as Pydoop and libraries such as SciPy) used in the ecosystem to
 1. **Visualize data** to make effective decision
 2. Understanding and **interpreting the value of customers' data** using scientific approach of data exploration
 3. Applying statistical methods with machine learning tools for predictive, descriptive and prescriptive **analytics**
- Apply **innovation** to integrated transportation network system (air, land, sea) and research ideas for continuous improvement by collecting more data **collection** via IOT sensors, mobile phones and collectors
- To achieve near **real time** processing of analytics using the resource of Hadoop, GPGPU or in memory servers.



Passenger Terminal Simulation: For an Intelligent and Integrated Transport System



With integrated data from air, buses, trains and vehicles, transportation planners could plan for upcoming smart terminals.



Passenger Terminal Analysis



Heatmap



During Peak Season in 2016, it is apparent that the capacity of the terminal being evaluated using the simulated parameters was over utilized by passengers. Therefore, can this terminal sustain without upgrading for the next 5 years with the evidence of historical data?



Conclusion



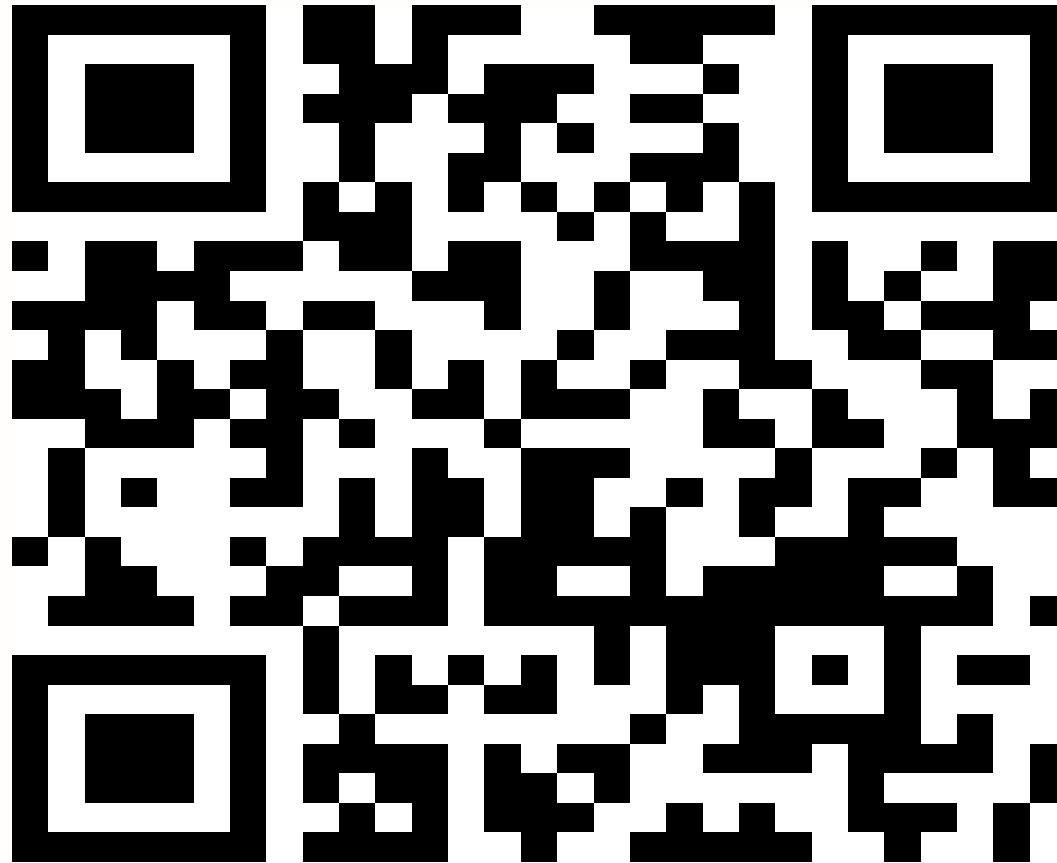
- Different type of solutions for big data analytics has been implemented by our team such as for transportation, public safety, health, social media analytics, video analysis, statistical analysis and for prediction of benchmarking datasets.
- Big Datasets in the order of hundred million to a billion rows of structured data are processed within Hadoop Distributed File System (HDFS), visualized, analyzed and further forecasted by users based on historical attributes
- It is necessary to analyze the datasets vs. the different configuration of resource allocations and techniques to be utilized such as when to use GPGPU, Hadoop, RDBMS and In-Memory solutions.
- We have filed patent for our resource allocations learning process (machine learning process) to efficiently utilize our Hadoop servers in our cloud.
- Integrated data from air and land further assists in the planning of smart passenger terminals and intelligent management of transportation system.



Question and Answer session



For more information
contact:
Dr. Keeratpal Singh
kps8822@gmail.com
+60193297900





Thank You



**Thank
You**