



Smart Geospatial Expo 2014

2014 스마트국토엑스포

International Conference on Geospatial Information Science 2014  
**Spatial Big Data Technologies and Applications For Future Society**  
미래사회를 향한 빅데이터 기술과 활용

2014. 08. 26 **coex** Room 317



Host |



**MOLIT**  
Ministry of Land,  
Infrastructure and Transport

Organizers |



**KRIHS**  
Korea Research Institute for  
Human Settlements



**LH**  
KOREA LAND & HOUSING  
CORPORATION



**KICT**  
KOREA INSTITUTE OF CIVIL ENGINEERING  
and BUILDING TECHNOLOGY







## 행사 세부일정 | Program

시간 (Time)		내용 (Title)	발표자 (Presenter)
09:00~10:00	60'	등록 (Registration)	
10:00~10:05	5'	개회사 (Opening Remarks)	손태락(국토교통부 주택토지실장) Tae-Rak Sohn (Head of Housing and Land Office, MOLIT)
10:05~10:10	5'	축사 (Congratulatory Address)	김경환(국토연구원장) Kyung-Hwan Kim(President, KRIHS)
10:10~10:40	30'	기조연설 (Keynote Speech): Adding Location to Big Data Analysis	레오나르드 제이아모한(ESRI 아시아 총괄이사) Leonard Jayamohan(General Manager, Public Sectors of ESRI Asia Pacific)
세션1: 정부3.0시대의 공간빅데이터 Session 1: Spatial Big Data – A New Resource for the Government 3.0 Era			
10:40~11:10	30'	정부3.0과 공간빅데이터 (Spatial Big Data For Government 3.0)	김대종(국토연구원 연구위원) Dae-Jong Kim(Research Fellow, KRIHS)
11:10~11:40	30'	공간빅데이터 개방, 공유, 연계 전략 (Linked Open Data – A Strategy for Sharing Spatial Big Data)	이경일(솔트룩스 대표) Tony Lee(CEO, Saltlux)
11:40~13:00	80'	점심 (Luncheon)	
세션2: 공간빅데이터 기술 Session2: Spatial Big Data Technologies			
13:00~13:30	30'	공간하둡: 공간빅데이터 분산처리 기술 (Spatial Hadoop: A MapReduce Framework for Spatial Big Data)	모하메드 모크벨(미국 미네소타주립대 교수) Mohamed F. Mokbel(Professor, Univ. of Minnesota, USA)
13:30~14:00	30'	비정형 빅데이터의 공간정보 처리 및 분석 기술(Spatial Analytics Platform for Unstructured Big Data)	정의선(한국오라클 상무) Eui-Seon Jung(Director of Oracle Korea)
14:00~14:30	30'	빅데이터 효율적 처리를 위한 클라우드 기술(REEF) 개발(REEF: Towards an Operating System for Big Data)	전병곤(서울대 교수) Byung-Gon Chun(Professor, SNU, Korea)
14:30~14:50	20'	휴식 (Coffee Break)	

※ 본 프로그램은 사정에 의해 일부 변경될 수 있습니다.

시간 (Time)	내용 (Title)	발표자 (Presenter)
세션3: 공간빅데이터 활용사례 Session3: Spatial Big Data Applications		
14:50~15:20	30'	미래 에너지와 도시 인프라를 위한 빅데이터 (Big Data for Future Energy and Urban Infrastructures – Challenges and Opportunities) 부드헨드라 바두리 (미국 오크리지 국립연구소 지리정보기술센터장) Budhendra Bhaduri(Director of Geographic Information Technologies Research Center, Oak Ridge National Lab, USA)
15:20~15:50	30'	국토모니터링을 위한 마이크로 지오데이터의 활용(Applications of Micro Geo Data for Urban Monitoring) 유키 아키야마(일본 동경대 연구위원) Yuki Akiyama(Research Fellow, Univ. of Tokyo, Japan)
15:50~16:20	30'	공간빅데이터를 활용한 행복주택사업 수요분석(How to Use Big Data in LH) 조연걸(한국토지주택공사 공간정보처 차장) Yeon-Gurl Cho(Vice Director of Spatial Information Division, Korea Land & Housing Corporation)
16:20~16:40	20'	휴식 (Coffee Break)
세션4: 패널 토론 Session4: Panel Discussion		
16:40~17:50	70'	주제: 공간빅데이터와 미래사회 Subject: Spatial Big Data and Future Society [좌장/Moderator] ▶오재인(한국빅데이터학회) Jay-In Oh(Korea Big Data Society) [토론자/Discussants] ▶손우준(국토교통부) Woo-Jun Sohn(MOLIT) ▶송규봉(GIS United) Kyu-Bong Song(GIS United) ▶성장환(토지주택연구원) Jang-Hwan Seong(LHI) ▶최현상(한국건설기술연구원) Hyun-Sang Choi(KICT) ▶황종성(한국정보화진흥원) Jong-Sung Hwang(NIA) ▶홍상기(한국공간정보학회) Sang-Ki Hong (Korea Spatial Information Society)
17:50~18:00	10'	폐회 (Closing)



## 해외 초청 연사 | Invited Foreign Speakers



 <p><b>Leonard Jayamohan</b> General Manager Esri Asia Pacific, Singapore</p>	<p><b>“Adding Location to Big Data Analytics”</b></p> <p>Leonard Jayamohan plays the role of General Manager for Esri in Asia Pacific. He is responsible for business development as well as relationship with key distributors and customers across the region. With more than 20 years of experience in Enterprise Business software, Leonard is quickly leveraging his knowledge of the enterprise business intelligence and applications and helping customers understand and adopt location analytics.</p>
 <p><b>Mohamed F. Mokbel</b> Professor University of Minnesota, USA</p>	<p><b>“Spatial Hadoop: A MapReduce Framework for Spatial Data”</b></p> <p>Mohamed F. Mokbel is an associate professor at University of Minnesota. His current research interests focus on providing database and platform support for spatio-temporal data, location based services 2.0, personalization, and recommender systems. Mohamed has held various visiting positions at Microsoft Research, USA, Hong Kong Polytechnic University, and as a founding Research Director of GIS Technology Innovation Center, Umm Al-Qura University, Saudi Arabia. Mohamed is an ACM and IEEE member and a founding member of ACM SIGSPATIAL.</p>
 <p><b>Budhendra Bhaduri</b> Corporate Research Fellow Oak Ridge National Laboratory, USA</p>	<p><b>“Big Data for Future Energy and Urban Infrastructures: Challenges and Opportunities”</b></p> <p>Dr. Budhendra Bhaduri is a Corporate Research Fellow and leads the Geographic Information Science &amp; Technology group at Oak Ridge National Laboratory. He also serves as the director of the Oak Ridge Urban Dynamics Institute. His research interests and experience include novel implementation of geospatial science and technology in sustainable development research, including human dimensions of critical infrastructure, urbanization and energy resource assessment.</p>
 <p><b>Yuki Akiyama</b> Research Fellow University of Tokyo, Japan</p>	<p><b>“Applications of Micro Geo Data for Urban Monitoring”</b></p> <p>Dr. Yuki Akiyama is a Research Fellow of the Earth Observation Data Integration and Fusion Research Institute of the University of Tokyo and a visiting researcher of the Center for Spatial Information Science of the University of Tokyo. He is chairman of Micro Geo Data Forum for the purpose of utilization and dissemination of “Micro Geo Data (MGD)” which is various kinds of big data with location information. His research interest focuses on monitoring, analysis and visualization of time-series urban changes using MGD.</p>



## 후원 | Sponsors



한국공간정보학회 Korea Spatial Information Society	한국지형공간정보학회 Korean Society for Geospatial Information System
한국지리정보학회 Korean Association of Geographic Information Studies	한국지도학회 Korean Cartographic Association
한국지리학회 Association of Korean Geographers	한국빅데이터학회 Korea Big Data Society



# Contents

I . Keynote Speech .....	1
* Adding Location to Big Data Analysis .....	3
II . [Session 1] Spatial Big Data – A New Resource for the Government 3.0 Era .....	13
1. Spatial Big Data For Government 3.0 .....	15
2. Linked Open Data – A Strategy for Sharing Spatial Big Data ...	33
III . [Session 2] Spatial Big Data Technologies .....	53
1. SpatialHadoop : A Map Reduce Framework for Spatial Big Data ...	55
2. Spatial Analytics Platform for Unstructured Big Data .....	62
3. REEF : Towards an Operating System for Big Data .....	79
IV . [Session 3] Spatial Big Data Applications .....	95
1. Big Data for Future Energy and Urban Infrastructures – Challenges and Opportunities .....	97
2. Applications of Micro Geo Data for Urban Monitoring .....	103
3. How to use Big Data in LH .....	117



# Keynote Speech

기조연설

Adding Location to Big Data Analysis  
: Leonard Jayamohan  
(General Manager, Public Sectors of  
Esri Asia Pacific)





## Adding Location to Big Data Analysis:

Leveraging Big Data's spatial information to derive insights and create repeatable information products

Leonard Jayamohan (General Manager, Esri Asia Pacific)

### 1. Background

Big Data has been the main topic of discussion in the analytics space. The huge amounts of data generated every day is creating the need to understand, assimilate and draw insights. This need is pushing the drive for analytics solutions. The power demonstrated by tools like Hadoop and SAP Hana which processes huge amounts of data in a very short time has spawned numerous projects and pilots to help organization derive actionable information to help in decision making.

So, what is Big Data? The definition of Big Data is that they are “sets of data that are too large and complex to manipulate or interrogate with standard methods or tools”<sup>1)</sup>. Let's try to understand what Big Data is from the perspective of governments or the public sector.



Figure 1 Types of Big Data

Some of the well-cited types of Big Data are weather, real-time traffic, demographics, consumer spending, income levels, crime, competition, labour force and sentiments. Take real-time traffic for example, a well sized city with about 2~3 million road vehicles would be generating more than 2~3 million strings of data in a second. Now that does not include the other traffic related data like number of commuters, accidents, construction blockages, etc. More on that later.

<sup>1)</sup> <http://business.wales.gov.uk/news-events/news/big-data-solution-launched>

In order to understand Big Data better, a lot of efforts has gone into categorizing them. These are known as the Three V's of Big Data: Volume, Velocity and Variety. I particularly like the categorization of the fourth V, Veracity and this is well illustrated in the following diagram, courtesy of IBM, see Figure 2 Four V's of Big Data.



Figure 2 Four V's of Big Data

The Four V's of Big Data are Volume, Velocity, Variety and Veracity. The definitions for these are:

- 1) Volume – Scale of Data. The sheer amount of data created and will be created due to the large number of connected devices increases. With more than 6 billion people with cell phones and a majority with smart phones, the amount of data generated by each user contributes to the sheer scale of data.
- 2) Velocity – Analysis of Streaming Data. A large amount of data is generated by devices or equipment in motion. Taking the cell phone user example, the location position and the direction of movement of the cell phones themselves at any point in time when analysed may yield insights into optimization of cellular services.
- 3) Variety – Different forms of Data. The sheer volume of data together with the value that can be derived by analysing that data in context with different data forms has allowed a lot of organization see trends or contextual value of the original data. In market analysis, demographic data that provides purchasing power, education level apart from age of consumers would give the any cellular service provider an idea of the types of services that a certain district or area might need.
- 4) Veracity – Uncertainty of Data. This is the hardest categorization of the Big Data. It refers to the biases, noise and abnormality in data<sup>2)</sup>. A good example is weather and its impact

2) <http://inside-bigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/>

of cellular services. How do we analyse that?

Standard methods and tools require some form of structured data and have limitations on the amount of data that can be crunched. These are the traditional business analytics tools. That is where high performance computing platforms and the proliferation of Hadoop helps. There is another tool that is beginning to be used by a number of government agencies to analyse Big Data and this is GIS tools<sup>3)</sup>. GIS software that has rich tool set to do spatial analysis provides the ability to layer the data (variety), map the frequency or density (volume), trace the movement (velocity) and identify the pattern over a period of time (veracity), gives new insights and supports better decision making.

In the Big Data and Cities event in Singapore<sup>4)</sup>, all the speakers were from the government agencies and in most of the presentations, we see the prevalence of GIS being used to understand the city better and then shape or provide services to better serve the city. From decisions on how to better build a road to service a new mixed development complex, stemming the spread of dengue by tracking the disease transmission, analysing the day time versus night time population in city centres, the speakers presented examples and results of analysis using GIS tools.

But before we look at GIS tools, let's take a step back to look at a subset of Big Data that has location components.

## 2. Location Based Big Data

In all of our activities, including sensor and machine activities, a large number of it is dependent on our location. Some would hazard a guess that in more than 80% of all activities we do, location or geography matters. These activities generate a large number of data and capturing these data with the location where the activity occurred can provide amazing insights.

When a city plans to serve the people, they need to figure out where and what kinds of services are required, how people can move and the transportation modes that can be provided. As urban areas become increasingly dense, cities and urban planners look at how to optimally place homes, businesses and workplaces and finally, where the urban dwellers can conduct recreational or shopping activities. These all are location based.

---

3) GIS refers to Geographic Information Systems.

4) <http://www.futuregov.asia/events/cities-big-data-summit-2014/>



Figure 3 People activities are Location Based

If we look beyond that, our sensor network is spread across a geographic area and its effectiveness is location dependent. A security camera is as good as the optical coverage it can provide, water height sensor is only cost effective if deployed over areas that may be prone to over-flooding.

### 3. GIS is in Transformation

GIS used to be the domain of the geographic analysts, and has been the tools that a large number of specialised analyst and their manager use for a long time. For them, putting data on a map and then using spatial analytics to derive patterns and insights have been their work for more than 30 years.

Today, GIS is in transformation. The work of these GIS Professionals now can bring value to many users as their analyses can be shared with executives using the browser. This is the emergence of Web GIS. Web GIS leverages the Big Data, other web services and the cloud by integrating volumes of data with Imagery, Lidar, sensor feeds and other forms of data to allow organizations to create content that is easy to disseminate to the masses. The end result is a thematic map that can immediately provide insight into the data.

In essence the GIS transformation allows us to integrate Geographic Science, the science of visual, into what we do, giving us different insights. I recently met with a police team who managed to lower crime by more than 20% over a period of 12 months. First,

they analyse the crime hotspots over a period of time spatially to create a pattern. Next, using that analysis, they deployed their patrol units according to these hotspots. A similar effort is shown in Figure 4 Crime Hotspots analysis for the San Francisco area.



Figure 4 Crime Hotspots analysis for the San Francisco area

## 4. Spatially analysing Big Data

Using GIS tools to analyse Big Data is not new. As the data sets become available and have location components to them, these can be analysed using GIS tools or with Big Data tools. Some of the examples show amazing results.

### 4.1 Big Data: Volume

Leveraging voluminous Big Data, some of the analysis has been to show sustainable land use, urban design and development. Apart from these, there were analyses that used social media (in this case Flickr). The volumes of data over a period of two years were mapped and thematically coded based on the country of origin of the visitors. What appeared was interesting enough to help the city decide where to provide Korean, Japanese and German translation assistance services for tourists.

Another example is leveraging topographic (read geographic features) of the landscape to plan for the cellular network. After setting up the network, the analysis continues using the signal spread of the radio base stations to help identify potential gaps in coverage.

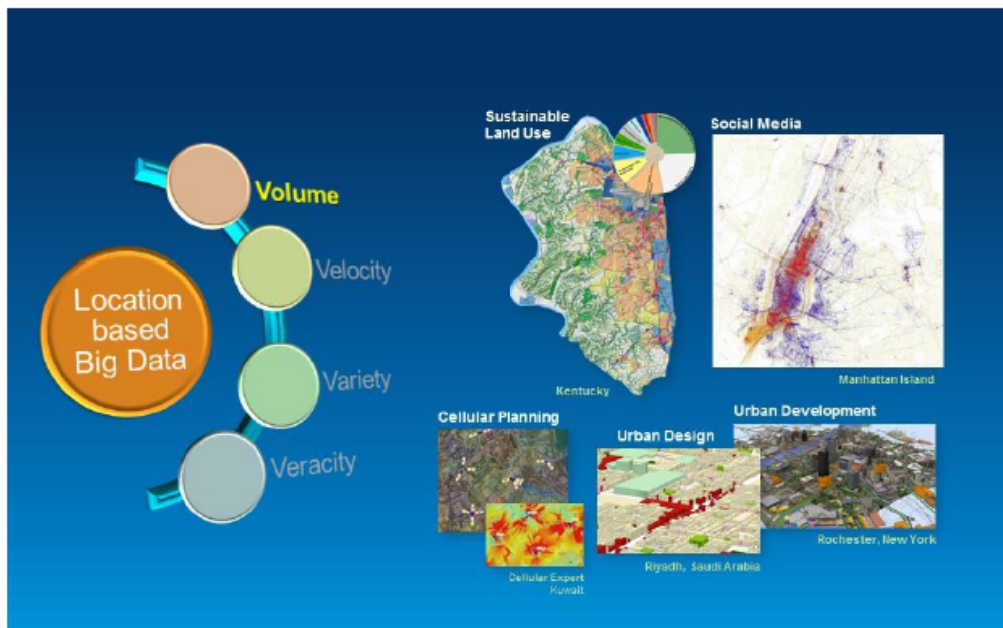


Figure 5 GIS analysis of Big Data – Volume

#### 4.2 Big Data: Velocity

One characteristics of Big Data is Velocity. This is a data set that is in a constant stream. Examples of analysis work done are real-time traffic, commuter movement or usage of public transportation and effectiveness of public transit.

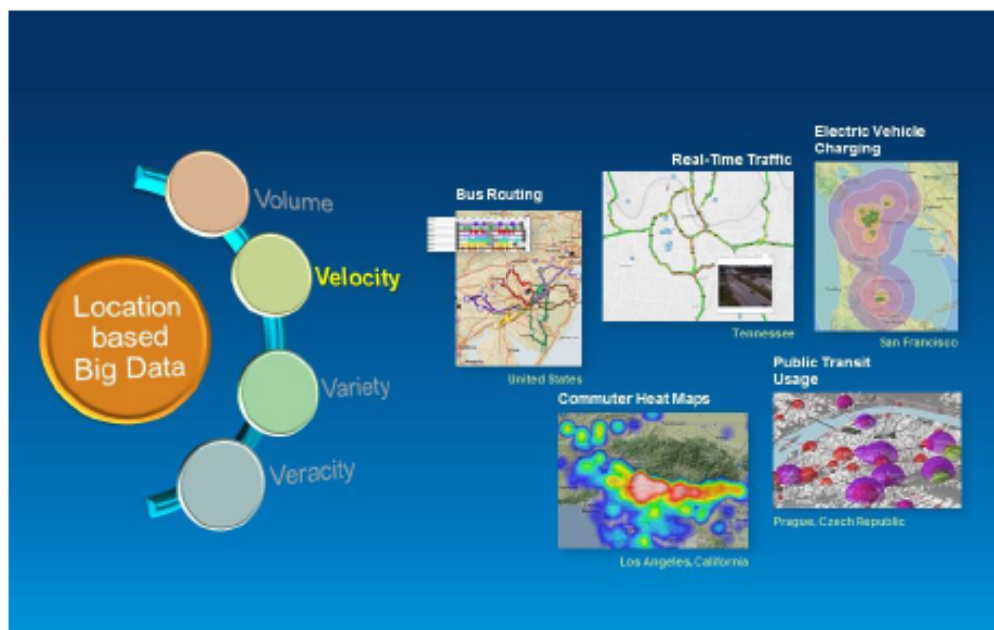


Figure 6 GIS Analysis of Big Data – Velocity



A very interesting development is the use of telematics data from vehicle GPS systems to analyse vehicle usage patterns to not only assist the road users get from one location to the next, but to advise the vehicle user on the next best time for servicing.

### 4.3 Big Data: Variety

GIS systems' ability to layer the data spatially and then analyse these layers to find patterns or trends, provides an interesting method to analyse data sets in relations to other data sets using location as a common reference point. An example from Singapore is the utilization of many layers from different agencies and real estate developers to plan and manage the use of limited land in Singapore.

Other examples are from the private sector where they used the profile of the land (buildings, road networks, malls, etc.) together with demographic information and their own network of dealers to identify gaps in dealer coverage, target addressable market and franchise planning.



Figure 7 GIS Analysis of Big Data-Variety

### 4.4 Big Data: Veracity

Due to the uncertainty of data like weather, a lot of the monitoring and tracking of inclement weather has been using GIS systems. A good example is the flood risk analysis which is dependent on many forms of data which is very hard to verify and validate. However, by taking all the data and looking at trends, an educated prediction based on the risk profile of the plains and its history of flooding can help organizations and individuals protect their assets.

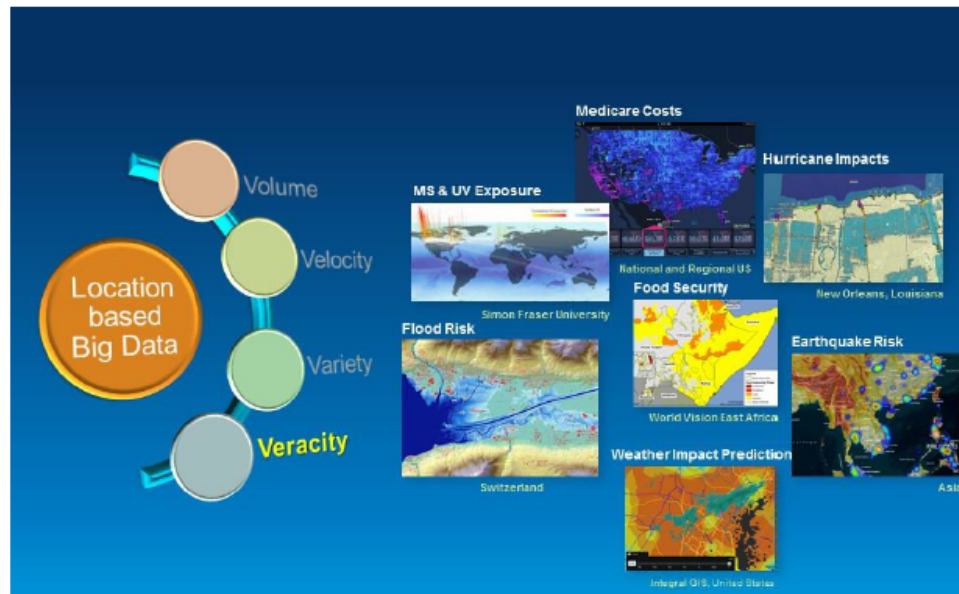


Figure 8 GIS Analysis of Big Data – Veracity

## 5. Case Study: Seoul Metropolitan Government

### 5.1 Floating Population Pattern analysis for Mid-Night Bus

SMG has run two night bus lines, which drive from midnight to 5 am, by the way of testing its effectiveness to enhance convenience for citizen. By the positive feedback of it, SMG decided to expand the number of bus lines up to 9 and, therefore, needed to confirm their routes. The team conducted the verification analysis using BIG DATA. It contains the anticipation of potential demand for midnight buses and modification of planned bus lines when necessary optional. 3 billion of mobile phone records for floating population pattern analysis and 5 million of taxi usage data for potential demand expectation were dealt with to confirm the best bus lines.

#### Project Purposes

- Collaborative work with commercial telecommunication company (KT) to support continuous increasing of demand for mid-night bus.
- Successful result of big data application in floating population analysis and examination of its capability on government issues.
- Suggest alternative routes based upon floating population data and SMG taxi usage data, which contain Original-Destination (OD) information.



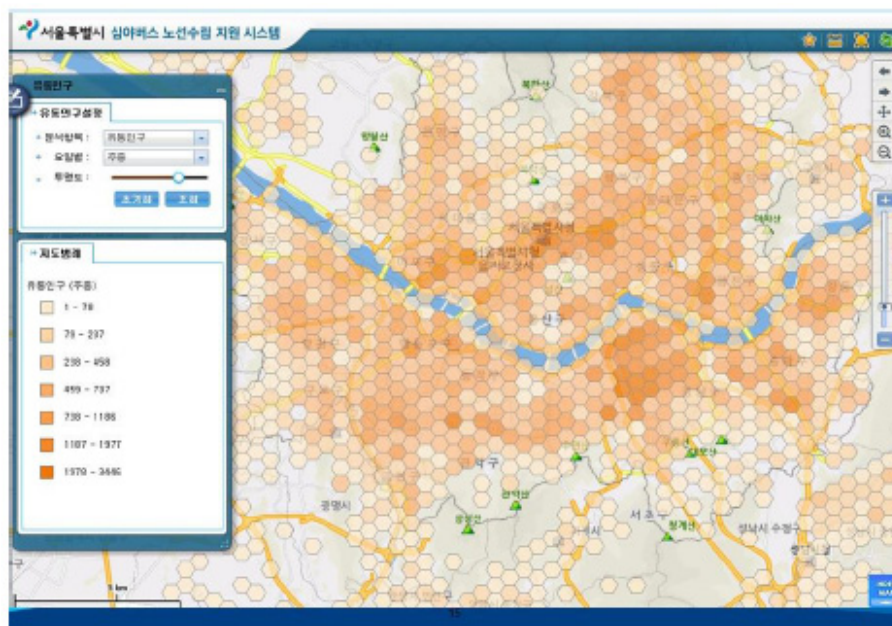


Figure 9 Web Application Example :  
Daytime Floating Population Distribution Status

#### Performance Details

- Building up floating population analysis model
- Verification of existing bus lines by applying assessment algorithm
- Visualization of population data per line/day to assess transit intervals
- Inquiry of OD direction information per zone and its mapping
- Creation of a report tool for statistical table such as the ranking list of demand per administrative DONG unit

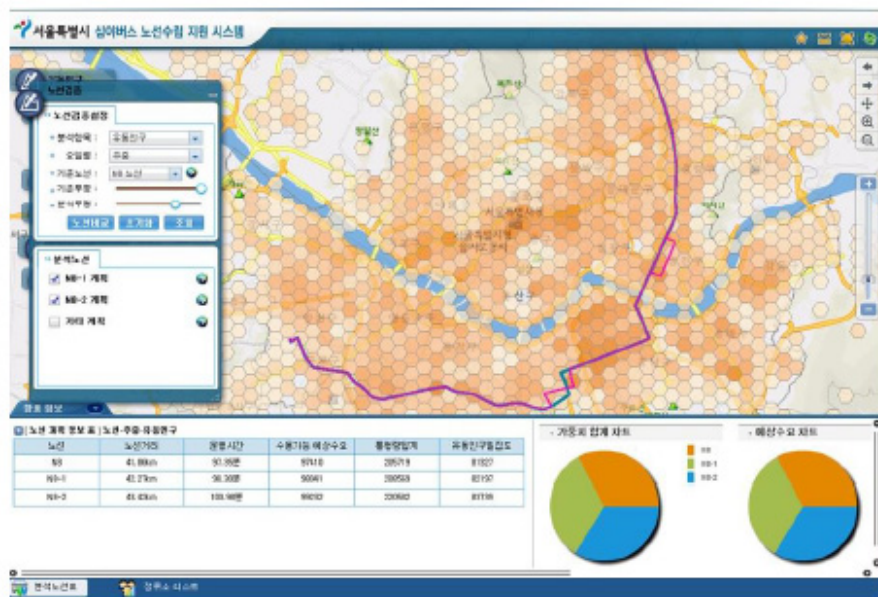


Figure 10 Web Application Example :  
Weighted Comparison and Evaluation by Bus Route

## 6. What's Next: GIS and Big Data

GIS is moving from 2D to 3D. In 3D, the representation of the actual locations is able to provide more insights into highly dense cities like Seoul, Tokyo, Beijing, Shanghai, etc. An interesting development is the use of 3D to better understand how to plan for a smarter and sustainable cities.

In Asia, the Urban Redevelopment Authority, is using 3D for urban planning, integrating many data sets, i.e. land profile, surrounding building mix, urban densities, demographic profile and supporting infrastructure like roads, power and water to plan new buildings and redevelopment of existing buildings.

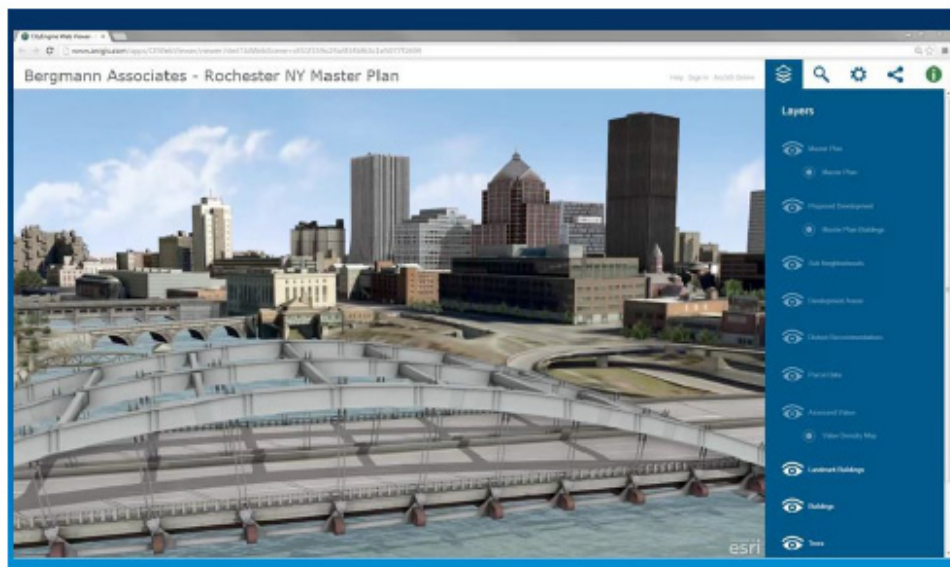


Figure 11 Planning, Modelling and Analysing Cities in 3D

# Session 1

정부3.0시대의 공간빅데이터

Spatial Big Data –  
A New Resource for the Government 3.0 Era

1. Spatial Big Data For Government 3.0  
: Dae-Jong Kim(Research Fellow, KRIHS)
2. Linked Open Data  
– A Strategy for Sharing Spatial Big Data  
: Tony Lee(CEO, Saltlux)



## 1

## Spatial Big Data For Government 3.0

Dae-Jong Kim (Research Fellow, KRIHS)

### ABSTRACT

The term of government 3.0 for Korea government was coined from the internet technology evolution and national agenda. Gov 3.0 is intrinsically based on world wide web 3.0 represented as semantic technology allowing customized and personalized information service. Strategies to achieve government 3.0 are openness and sharing data by government, communication between people and government, and collaboration among government parties.

Big Data is emerged as an avenue for achieving government 3.0 because it gives chance to understand the real world better and to make better decision based on insight. Especially, spatial Big Data allows detailed and rich insight since behavior and planning of people in big data can be illustrated on map and interpreted in the spatial context. Spatial Big Data is also very effective in communication. It has been told that about 80% of big data is geographically referenced.

Two case studies were illustrated in the presentation. One is to predict land use change by detecting spatiotemporal patterns of land transaction data. It turned out that detected patterns on agricultural land and forest were converted into urban land use several years later. The other is to diagnose current issues in real estate market and evaluate the effect of policies. Opinion and sensitivity analyses using SNS data were effective to understand what people are saying about policies. Spatiotemporal pattern analysis of contract data for rental housing and housing transaction data revealed that tax reduction/exemption policy induced rental housing demand to purchase but didn't stop sky-rocketing rental housing price.

Some premises are identified for successful Big Data based government 3.0. First, strong institutional basis for making big data being produced in the public business process PUBLIC and for dealing with privacy and anonymization. Second, easy access and application to Spatial Big Data is essential. Platform service including Spatial Big Data, methods and tools is urgent. Lastly, technology development such as in-memory DBMS and HW accelerated hadoop platform are required to process spatial Big Data for policy making just in time and in place.





[International Conference on Geospatial Information Science 2014]

# Spatial Big Data for Government 3.0

2014. 8



Kim, Daejong(Ph.D)



KRIHS 국토연구원

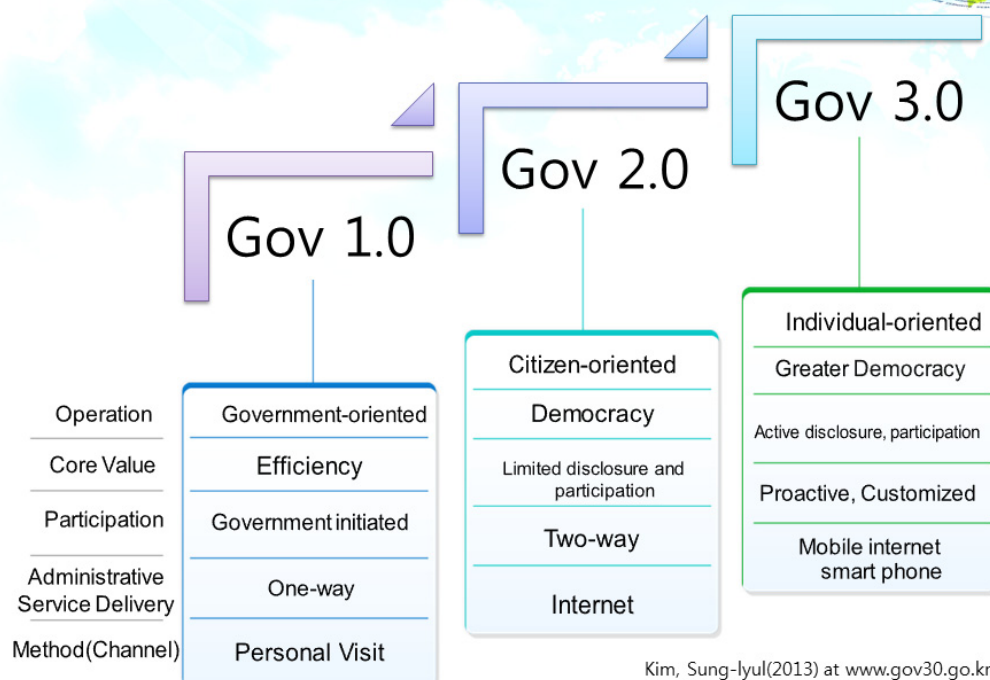


## Contents

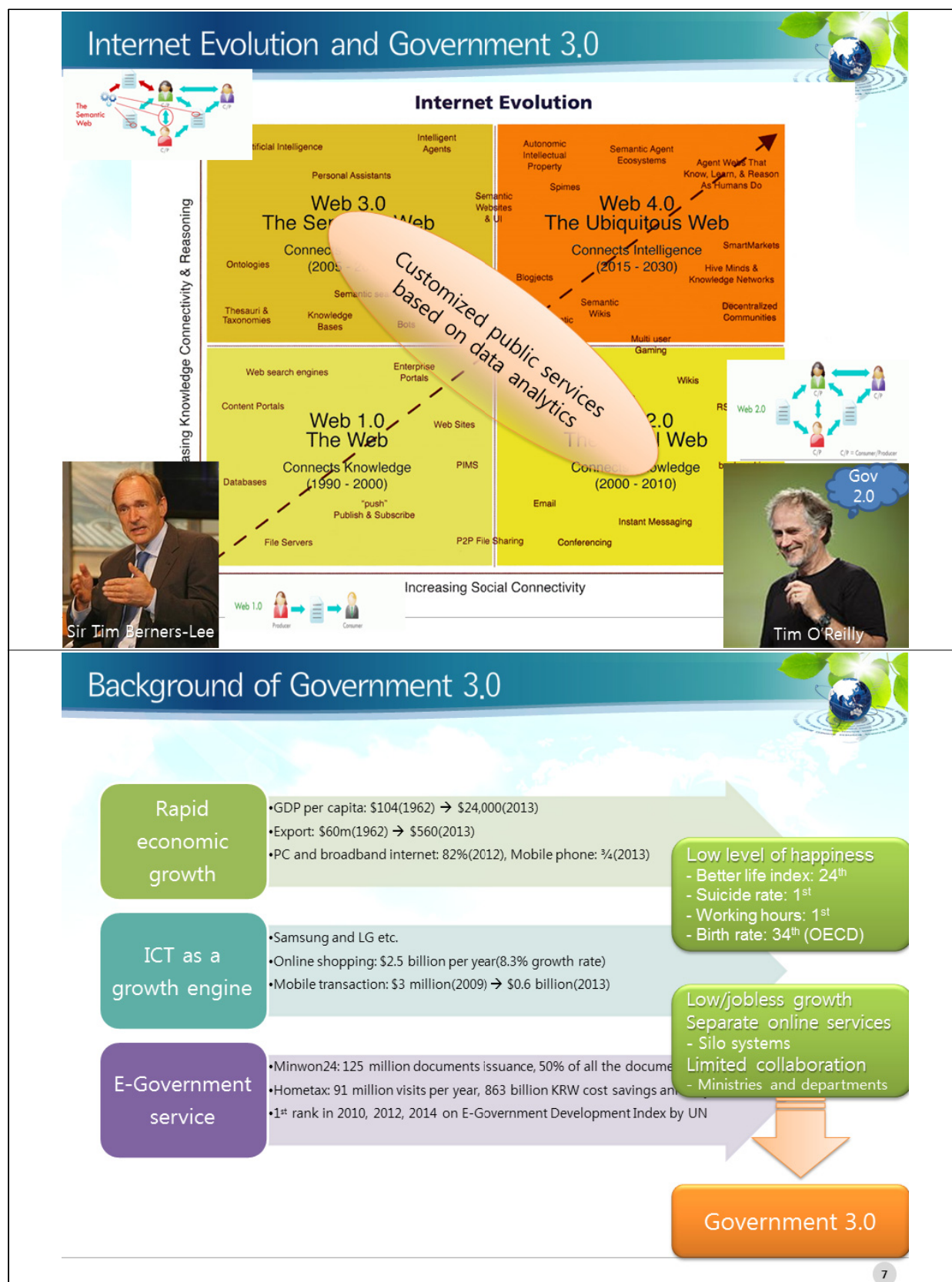
- I. What is Government 3.0?
- II. What is Spatial Big Data?
- III. Case Studies
- IV. Premises for Successful Gov 3.0



## Government Evolution



Kim, Sung-lyul(2013) at [www.gov30.go.kr](http://www.gov30.go.kr)



**Rapid economic growth**

- GDP per capita: \$104(1962) → \$24,000(2013)
- Export: \$60m(1962) → \$560(2013)
- PC and broadband internet: 82%(2012), Mobile phone: ¾(2013)

**ICT as a growth engine**

- Samsung and LG etc.
- Online shopping: \$2.5 billion per year(8.3% growth rate)
- Mobile transaction: \$3 million(2009) → \$0.6 billion(2013)

**E-Government service**

- Minwon24: 125 million documents issuance, 50% of all the documents
- Hometax: 91 million visits per year, 863 billion KRW cost savings and
- 1<sup>st</sup> rank in 2010, 2012, 2014 on E-Government Development Index by UN

**Low level of happiness**

- Better life index: 24<sup>th</sup>
- Suicide rate: 1<sup>st</sup>
- Working hours: 1<sup>st</sup>
- Birth rate: 34<sup>th</sup> (OECD)

**Low/jobless growth**  
**Separate online services**

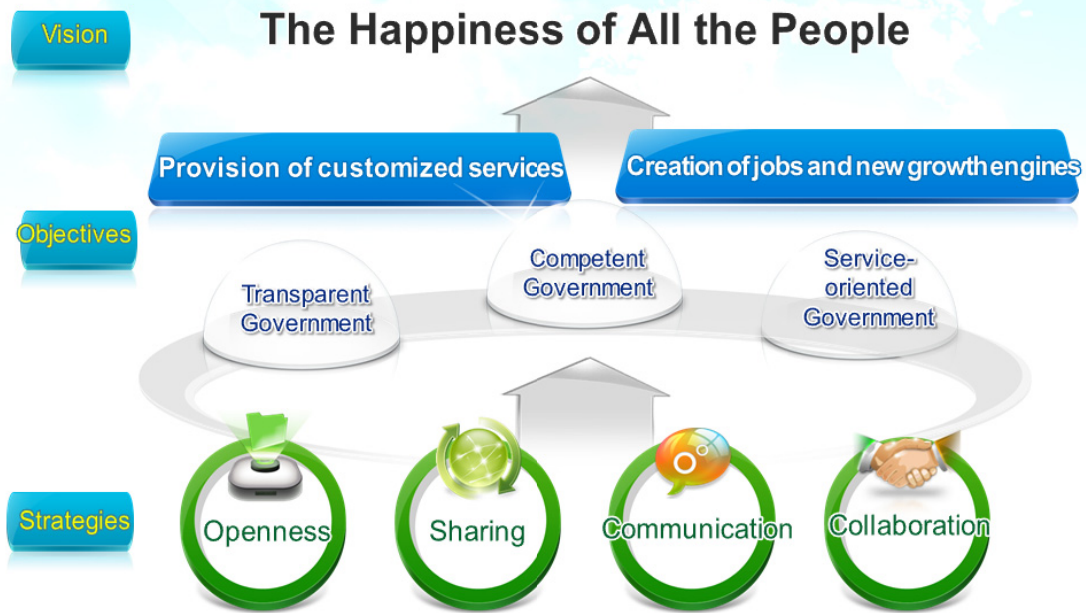
- Silo systems
- Limited collaboration
- Ministries and departments

**Government 3.0**

7



## Vision and Strategies of Government 3.0



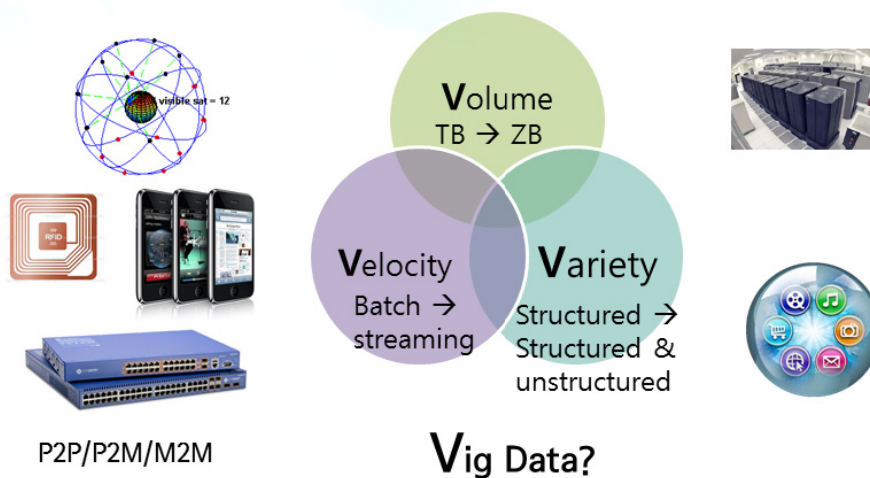
Kim, Sung-lyul(2013) at [www.gov30.go.kr](http://www.gov30.go.kr)

8



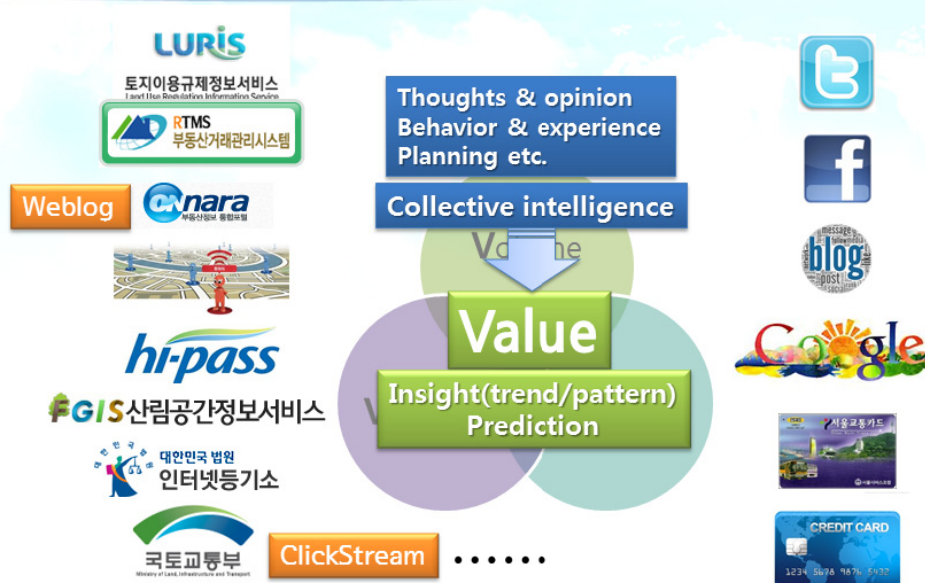
## What is Big Data?

“a collection of **data sets** so large and complex that it becomes **difficult to process using on-hand database management tools or traditional data processing application.**” - Wikipedia, 2014 -



10

## Why is Big Data?

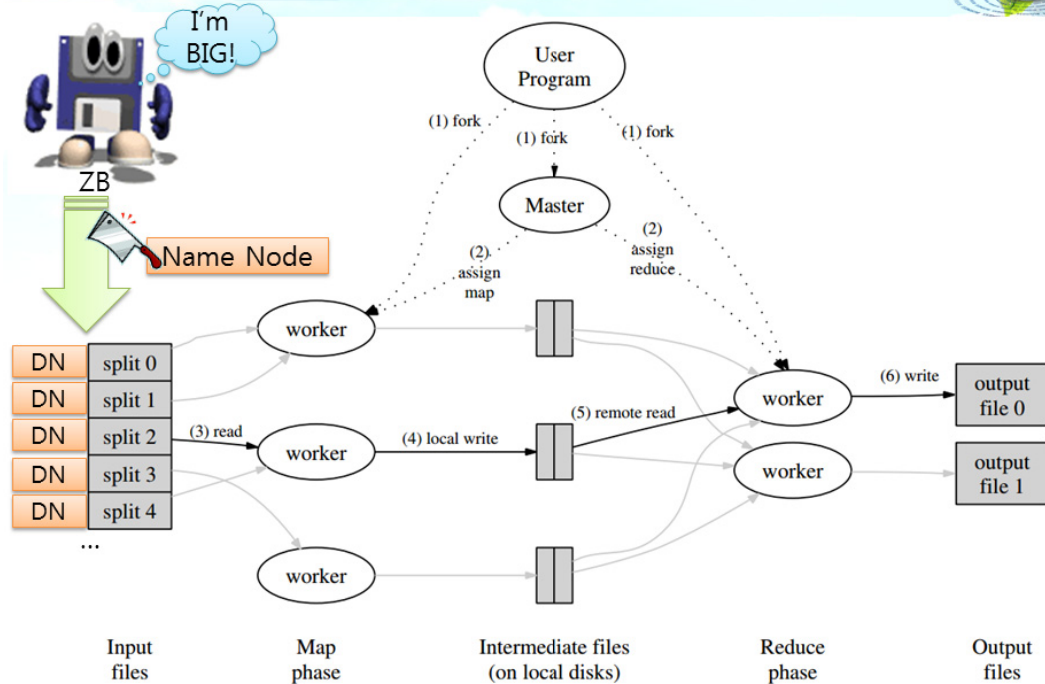


“shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions.”

-New York Times-

11

## Collaboration System for Big Data



Source: Jeffrey Dean and Sanjay Ghemawat (Google, Inc.). 2004

12

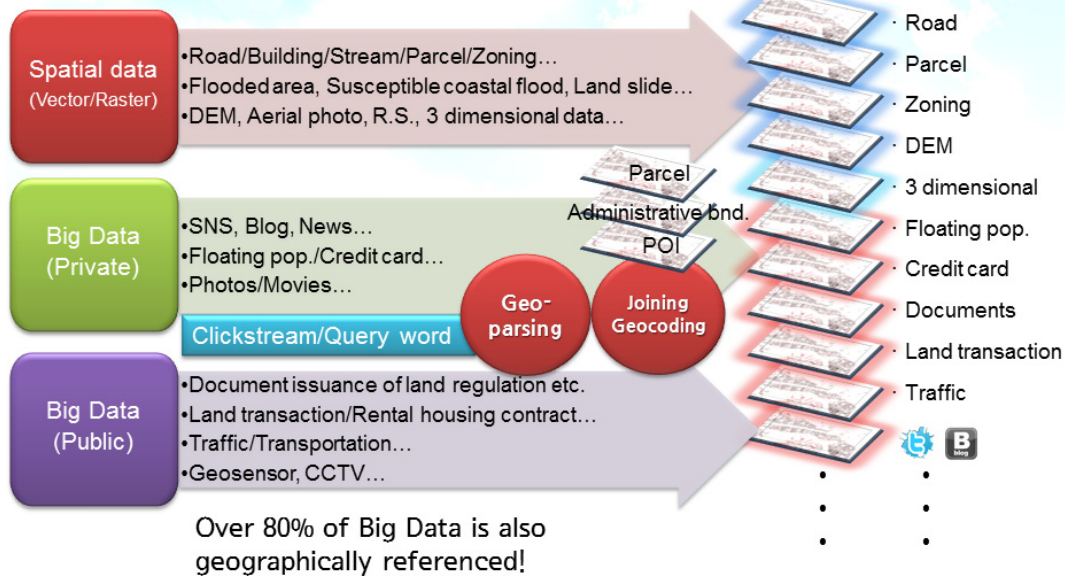
## Hadoop Ecosystem



13



## What is Spatial Big Data?

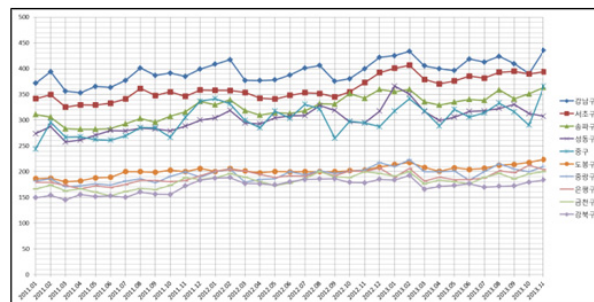


14

## Why is Spatial Big Data?

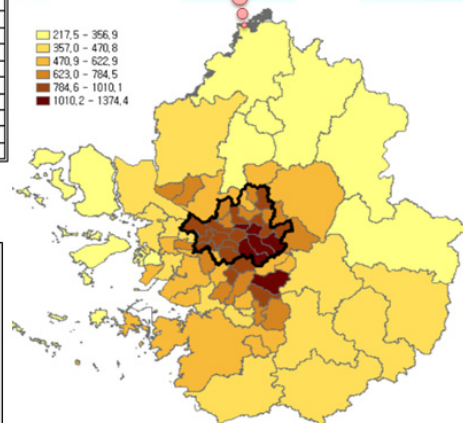
	2011.01	2011.02	2011.03	2011.04	2011.05	2011.06	2011.07	2011.08	2011.09	2011.10
강남구	371.4	393.9	356.4	352.9	365.4	363.4	377.0	401.1	386.8	391.6
서초구	341.6	349.6	325.4	329.6	329.3	332.9	341.1	361.6	347.8	
송파구	311.2	306.0	283.5	282.4	282.5	283.5	292.9	303.4	296.3	307.0
성동구	273.8	287.9	257.5	261.3	270.9	280.0	279.1	284.5	282.8	279.0
동구	243.9	296.0	266.7	267.7	262.5	261.2	269.2	285.3	285.1	266.6
마포구	255.3	264.6	265.4	256.7	253.3	269.2	279.9	278.7	271.1	271.6
동작구	260.7	269.2	257.7	248.6	255.2	255.1	255.4	272.1	267.9	256.8
영등포구	252.4	265.1	256.6	258.3	258.9	262.1	258.5	262.9	270.8	271.6
양천구	296.9	314.3	267.9	251.8	247.6	259.0	254.1	263.0	242.5	248.5
관악구	259.0	266.8	236.3	231.2	228.7	237.2	248.5	263.2	260.9	245.4
부천시	231.5	257.7	250.0	247.8	228.8	227.2	237.4	256.9	236.7	243.0
중구	226.5	234.3	219.7	231.6	227.3	216.0	235.0	229.0	238.8	224.2
관악구	238.1	252.5	235.7	225.5	220.0	230.8	240.2	246.1	240.0	233.3
강서구	224.1	234.3	211.9	219.3	217.8	220.7	219.3	231.5	227.3	225.0
강동구	224.0	212.0	201.6	210.9	212.3	212.4	218.8	229.2	232.2	235.3

Rent price (median value) per unit for each city



Rent price per unit for each city

Understanding and effective communication  
→ customized policy



Rent price map per unit for each city

15

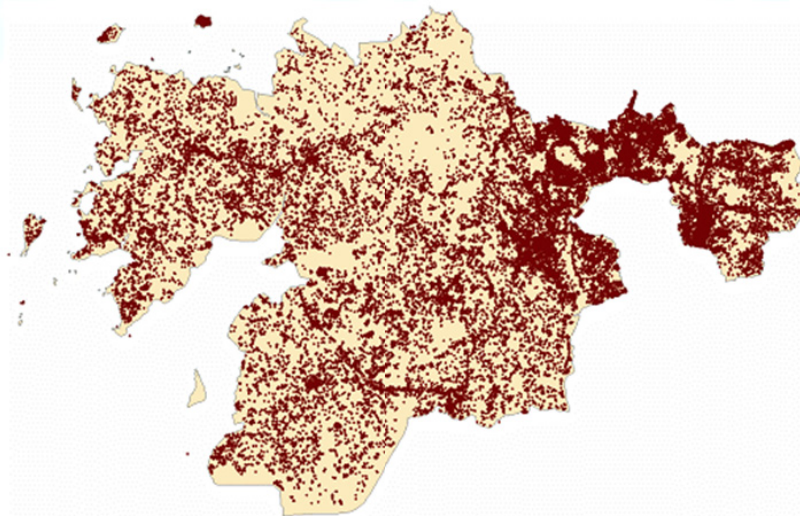
## III Case Studies

Land Use Change Prediction

Real Estate Market Monitoring

### Land Use Change Prediction

- ☐ H1: There is significant spatiotemporal clustering in land transactions where large-scale land development will occur
- ☐ Geocoded land transactions (328,855/436,804) for 2001-2010



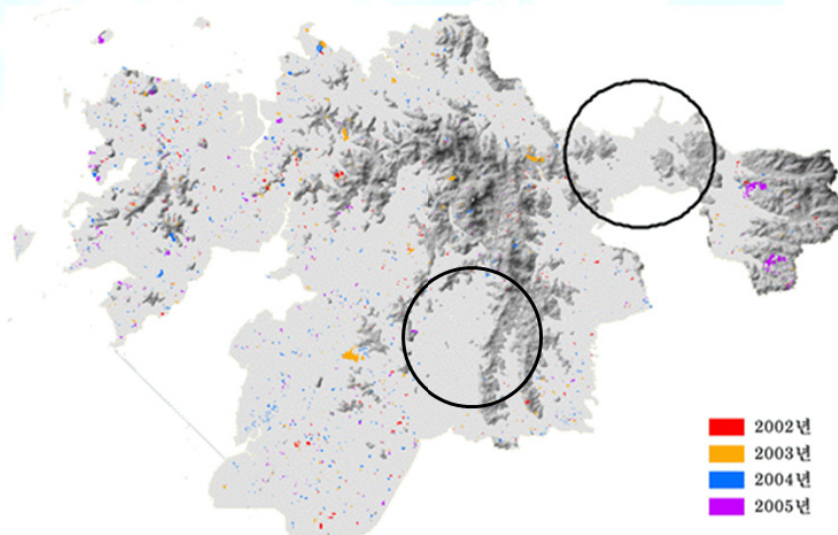
Source: Dae-jong Kim-Hyeong-su Koo. 2011. Land Use Change Prediction with Spatiotemporal Pattern Analysis and Strategies for Urban Policy

17



## Land Use Change Prediction

- Spatiotemporal patterns detected (2002-2005) using spatio-temporal chain statistics

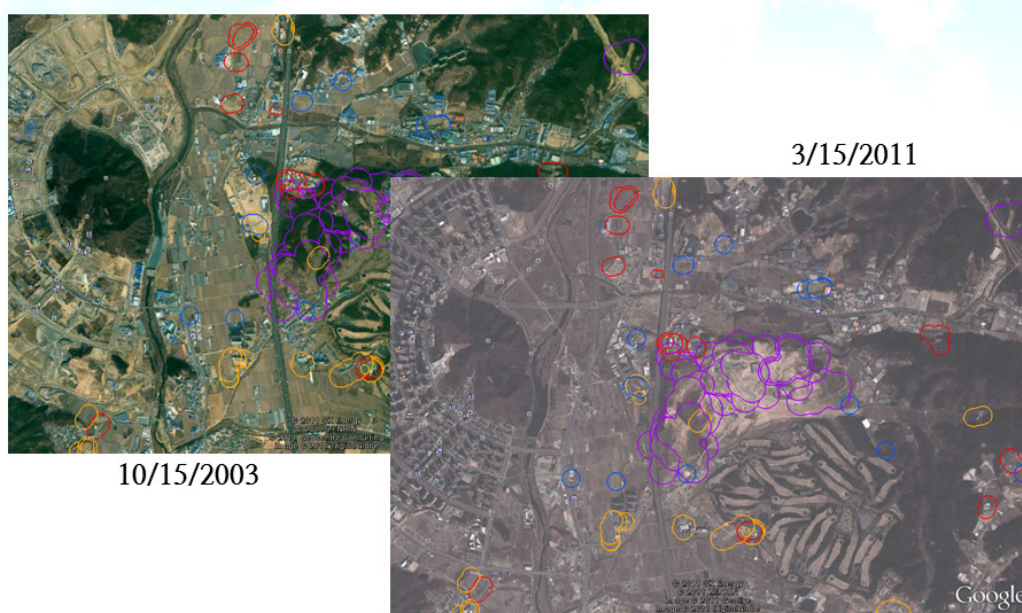


Source: Dae-jong Kim Hyeong-su Koo. 2011

18

## Land Use Change Prediction

- Visual evidence

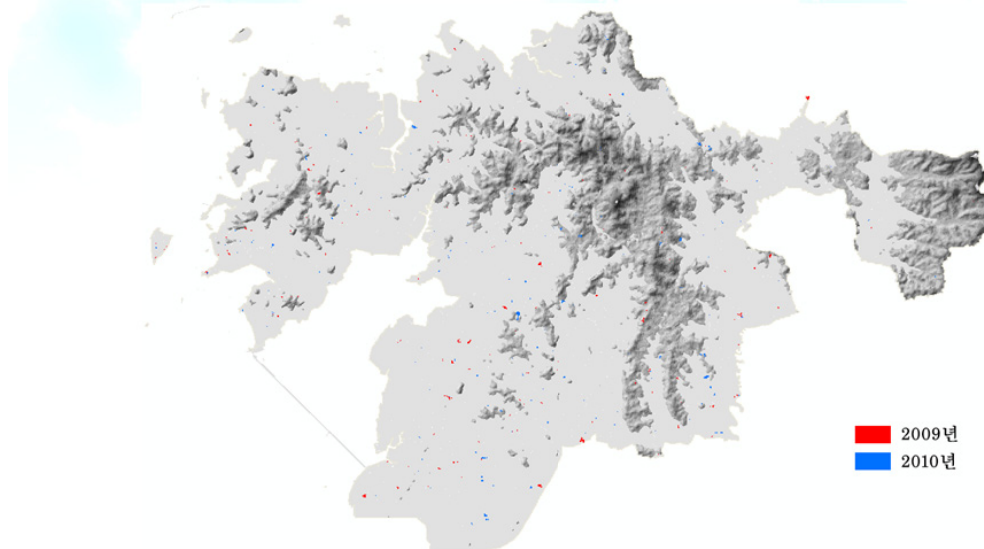


Source: Dae-jong Kim Hyeong-su Koo. 2011

19

## Land Use Change Prediction

- Detected spatiotemporal patterns of land transactions in 2009-2010



Source: Dae-jong Kim Hyeong-su Koo. 2011

20

## Land Use Change Prediction

- Detected spatiotemporal patterns in land transactions in 2009-2010 where land development is expected in the future



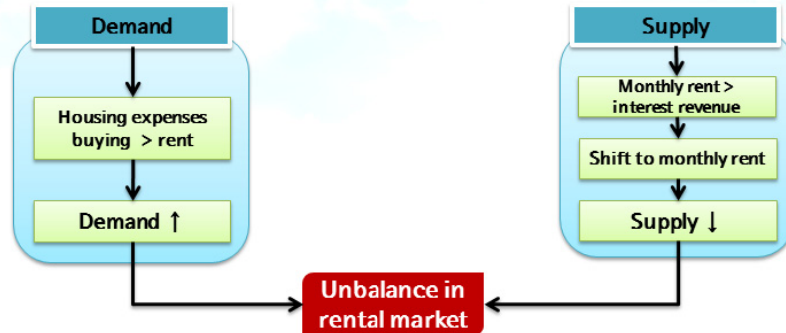
Source: Dae-jong Kim Hyeong-su Koo. 2011

21



## Real Estate Market Monitoring

- Issues: Unbalance between demand and supply in JEONSE (Deposit w/o monthly rent) market → Sky-rocketing rent expense in Seoul metropolitan area



- Policies
- Acquisition/transfer income tax reduction/exemption to induce rent demand to buying

22

## Real Estate Market Monitoring

- Opinion analysis(Frequently mentioned words)
- Keywords: House, apartment, villa, real estate etc.
  - Collection: Twitter, café, blog, news
  - Period: 2013. 3 – 2013. 9
  - Keywords: 'rent expense', 'loan', 'rent shortage', 'rent poor', 'tin rent'

Mo.	3	4	5	6	7	8	9	
Rank	Keyword	Freq	Keyword	Freq	Keyword	Freq	Keyword	Freq
1	전세자금	4	전세가격	778	전세	918	전세	1,059
2	전세가격	3	전세	707	전세가격	785	전세가격	868
3	전세	2	전세자금	479	전세자금대출	275	전세시장	482
4	전세탈출	1	전세자금대출	273	전세가비	260	전세난	248
5			전세시장	132	전세시장	90	전세자금	211
6	역전세난		전세수요	89	전세수요	85	전세자금대출	278
7	서민전세자금		전세자금	82	반전세	83	전세가비	204
8	전세수급지수		장기전세	63	목돈안드는전세제도	67	근로자서민전세자금	131
9	전세가비		전세계약	55	전세임대	23	전세제도	94
10	전세제도		전세임대	45	장기전세	34	전세수요	82
11	반전세		전세난	42	전세물건	33	전세계약	58
12	매매전세물세		전세물량	39	전세자금대출	30	전세물건	24
13	전세계약		전세비중	34	서민전세자금	28	전세가비	19
14	전세임대		전세가격지수	33	전세물량	24	전세임대주택	16
15	장기전세		전세계약	32	전세수요	15	전세자금증역	15
16	전세계약기간		장동전세	28	전세매물	13	전세잔액	15
17	전세물건		전세가격변동률	18	전세제도	7	매매전세물세	13
18	전세상물		전세제도	9	전세비중	6	전세매물	12
19	전세난		전세가구	12			전세보증금반환보증	11
20	전세물량		반전세	9			전세자금대출금리	10
21	전세지수		전세물건	8			전세살이	9
22	신혼부부전세자금			7			전세아파트	7
23							전세대출금	7

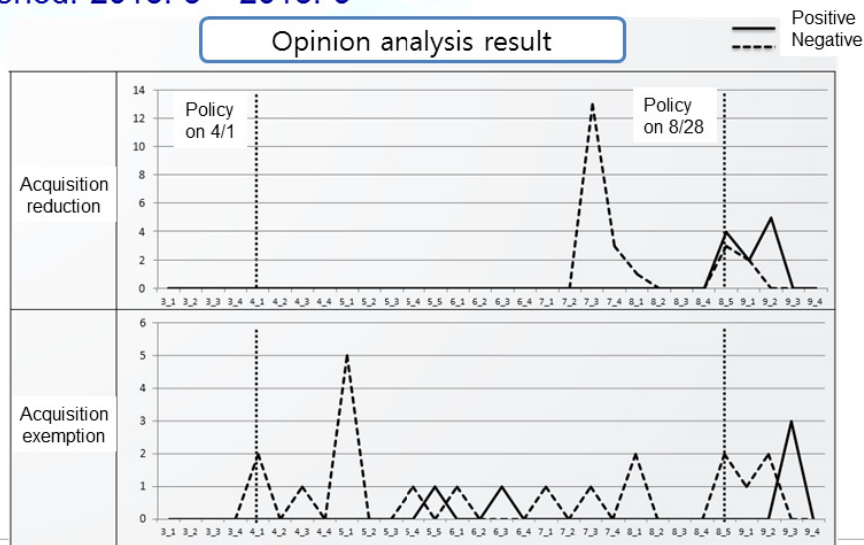
23



## Real Estate Market Monitoring

### □ Opinion/sentimental analysis

- keyword: Real estate policies on April 1<sup>st</sup> and on August 28<sup>th</sup>
- Collection: Twitter, café, blog, news (93,877 pages)
- Period: 2013. 3 – 2013. 9



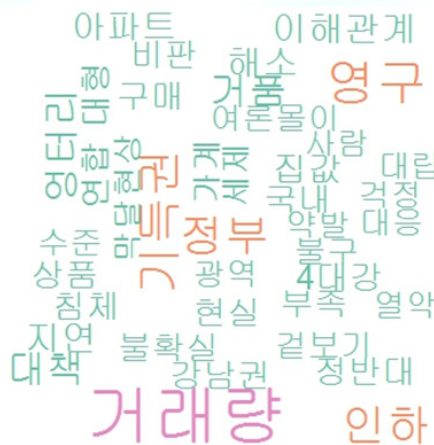
24

## Real Estate Market Monitoring

### □ Related issue analysis

- Keywords from negative responses
- 'Temporal policy', 'vested rights', 'bubble', 'last month'

#### Word cloud

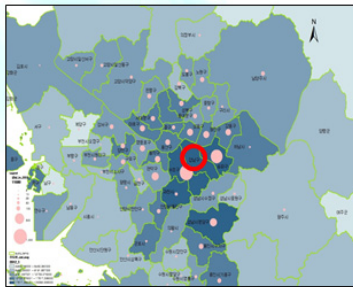


Word	Freq.
거래량	186
영구	107
기득권	94
정부	79
인하	74
거품	48
대책	46
영터리	45
막달현상	32
일시적	32
한심	31
국책	30
이해관계	30

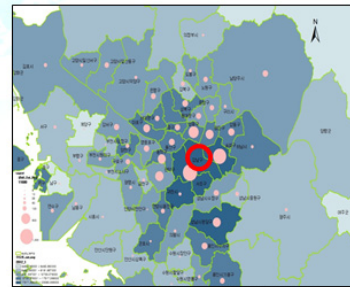
25

## Real Estate Market Monitoring

### □ Social network analysis

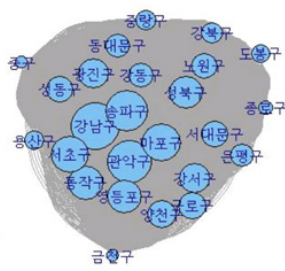


<Gangnam-gu: Influx>

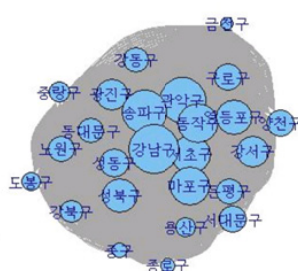


<Gangnam-gu: Efflux>

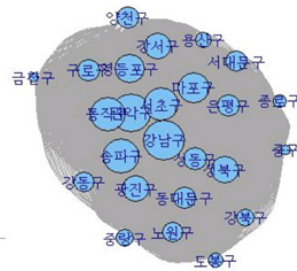
Degree Centrality(2011)



Degree Centrality(2012)



Degree Centrality(2013)

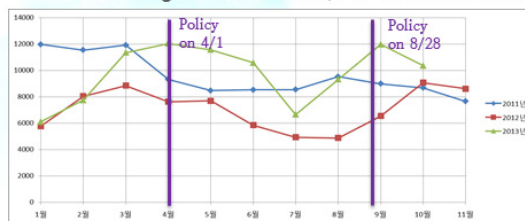


26

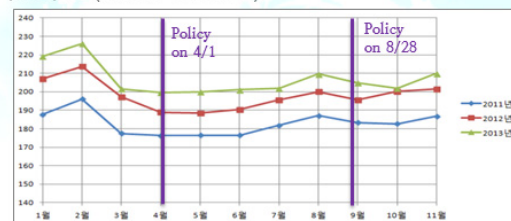
## Real Estate Market Monitoring

### □ Policy effect analysis

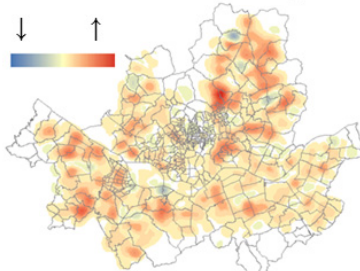
- Housing transaction 306,140 Rental contract: 2,789,662(2011/1- 2013/11)



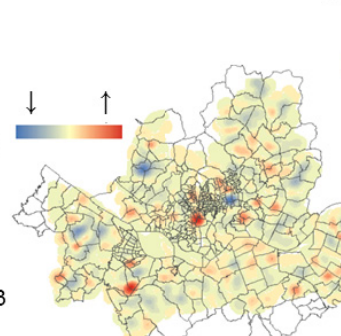
Amount of housing transaction



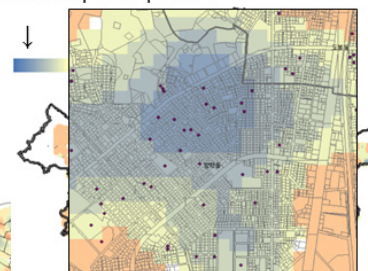
Rental expense per unit



Amount of housing transaction in 2013 compared to 2012



Change rate of housing price in 2013 compared to 2012



Change rate of rental price in 2013 compared to 2012

27

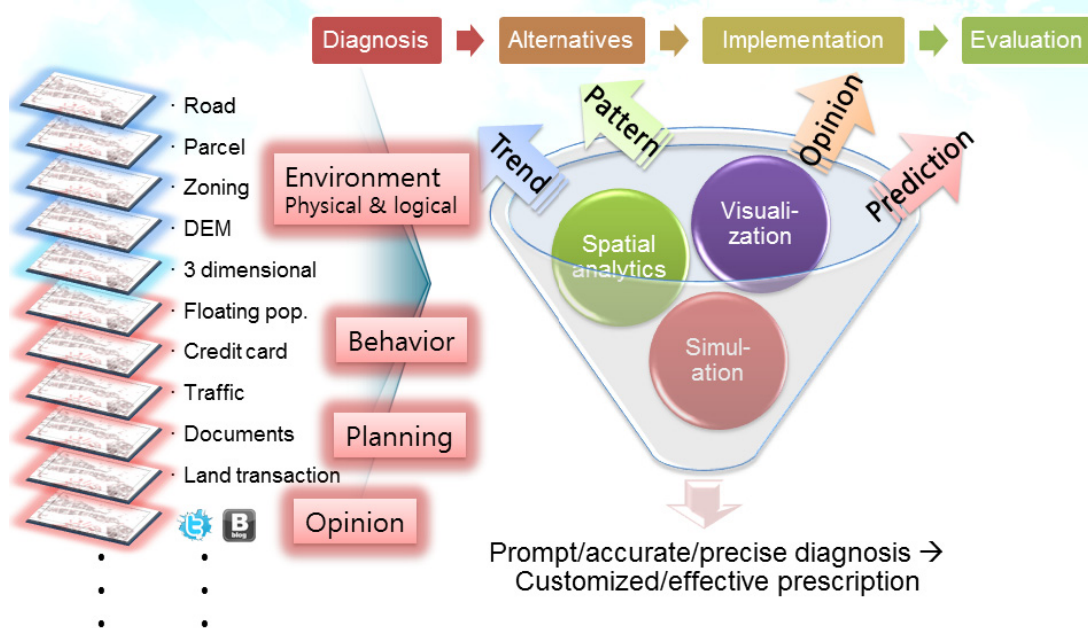
## IV Future Directions for Gov. 3.0

Application Model of Spatial Big Data

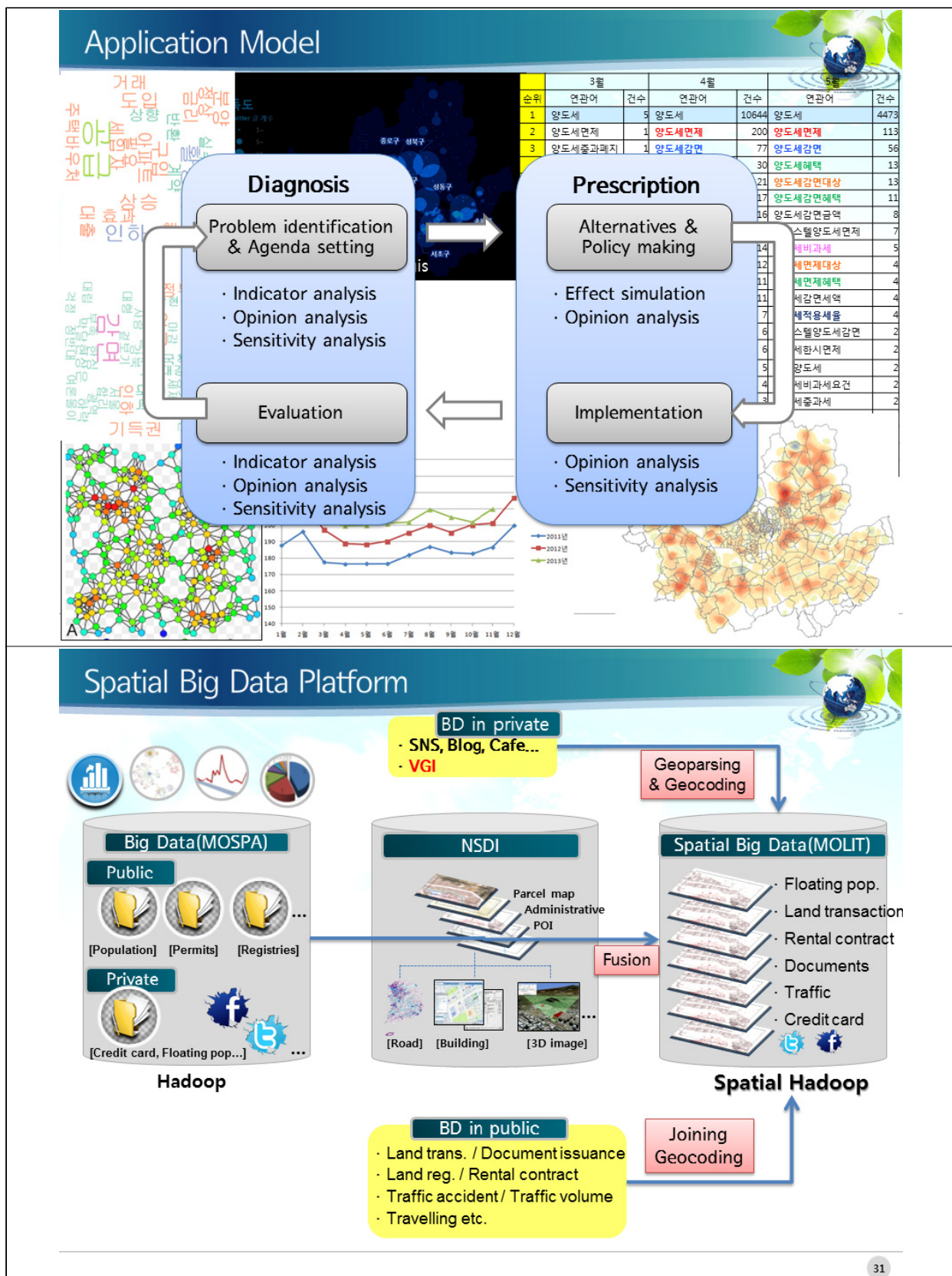
Spatial Big Data Platform Development

Premises for Successful Gov. 3.0

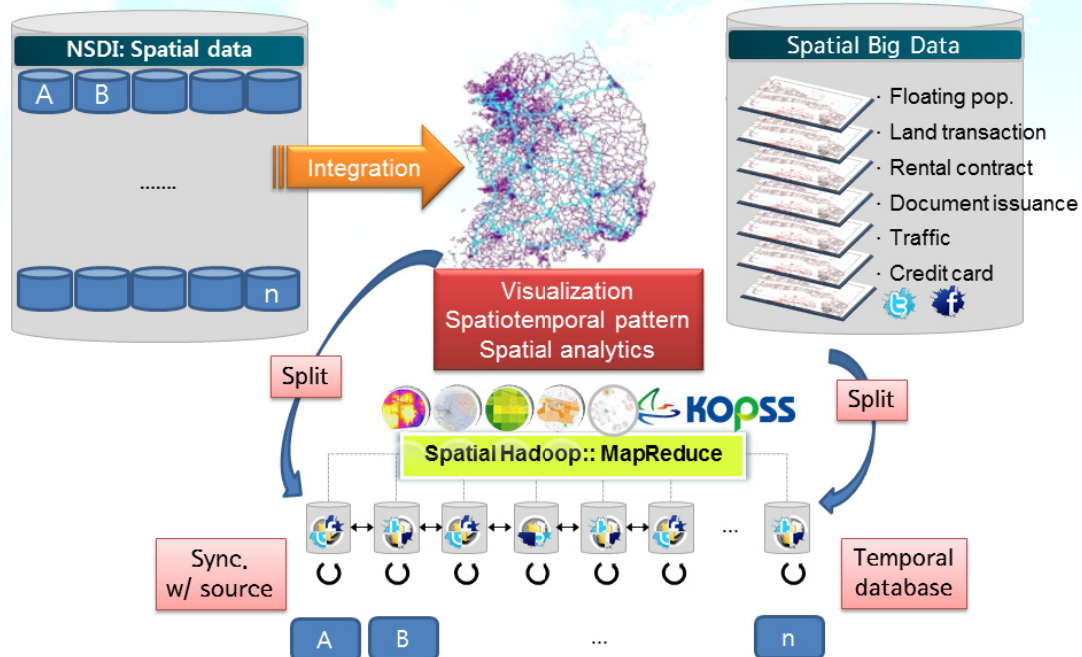
### Application Model





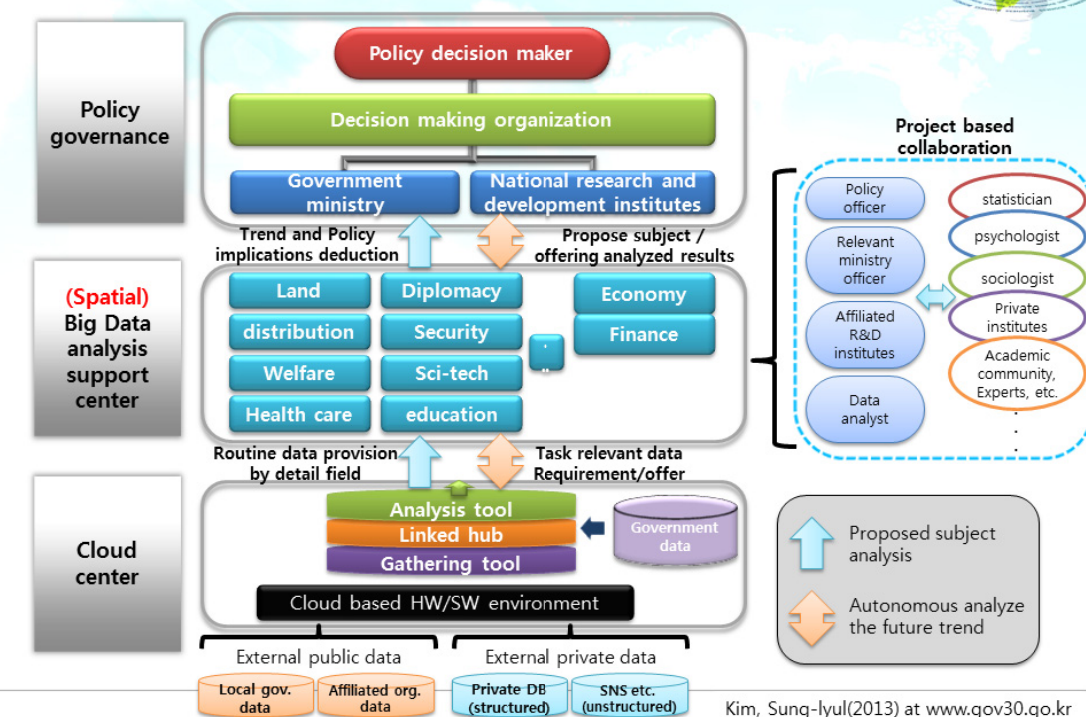


## Spatial Big Data Platform



32

## National Future Strategy Center



## Premises for successful Big Data based Gov. 3.0



### Strong institutional basis for opening public big data

- Making Big Data being produced in the public business process PUBLIC
- Privacy and anonymization

### Platform service for utilization

- Not only Spatial Big Data but also methods/tools for utilization
- Easy access and participation

### Technology development for computational infrastructure

- In-memory DBMS, HW accelerated
- Analytical methodologies

34

# Thank you!

Daejong Kim (Ph.D)  
djkim@krihs.re.kr  
Geospatial Research Division  
Korea Research Institute for Human Settlements

## 2

## Linked Open Data - A Strategy for Sharing Spatial Big Data

Tony Lee (CEO, Saltlux)

ICGIS 2014 - Seoul Coex, 26th Aug.

# Linked Open Data

## A Strategy for Sharing Spatial Big Data

Tony LEE / [tony@saltlux.com](mailto:tony@saltlux.com)

CEO and President of Saltlux, Inc.



Tony LEE

Saltlux, Inc.

CEO and President

- Exec. Advisor for Korean Gov.
  - Big data advisory board of MSIP
  - Government 3.0 advisory board of MOSPA
  - Committee member of Open Data Council
- Inha Univ. , Adjunct Associate Prof.
- KM/EDM Association, Chairman
- ISO TC37, Committee Member
- Society for CI, Board Member
- STI International, Board Member
- KICT, Honorary Researcher
- LG Central Lab., Researcher
- Association for HCI, Chairman
- Chairmen for ISWC and etc.



# Government 3.0 and Open Data

3

## Government 3.0 in Korea

### Rebooting government

**G**overnment will make 100 million information disclosures annually, up from 310,000 last year. Full contents will be made public, and the government vows to classify a minimum amount of data as confidential.

**G**overnment will offer its data for commercial use to boost entrepreneurship and make a "creative economy."

**G**overnment will collect opinions on major policies and projects and seek cooperation with the private sector through online, direct democracy.

Source: Ministry of Security and Public Administration



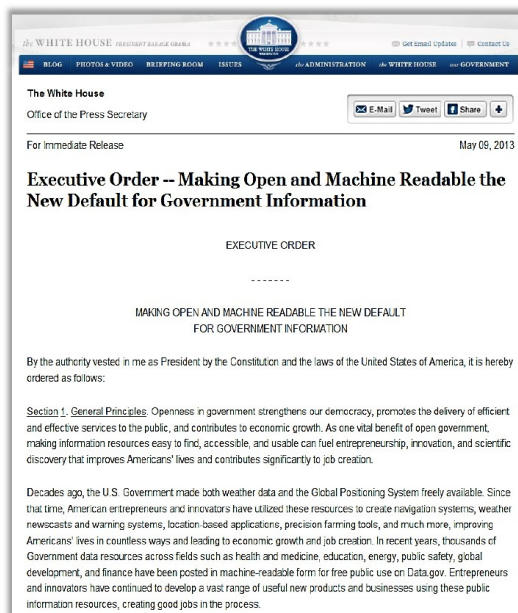
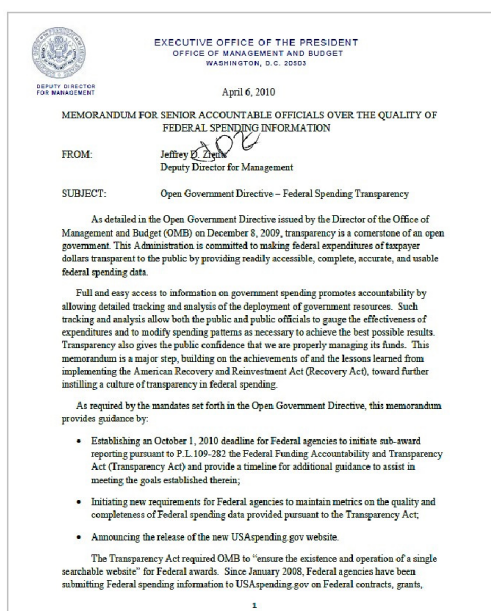
4

## Government 3.0 in Korea



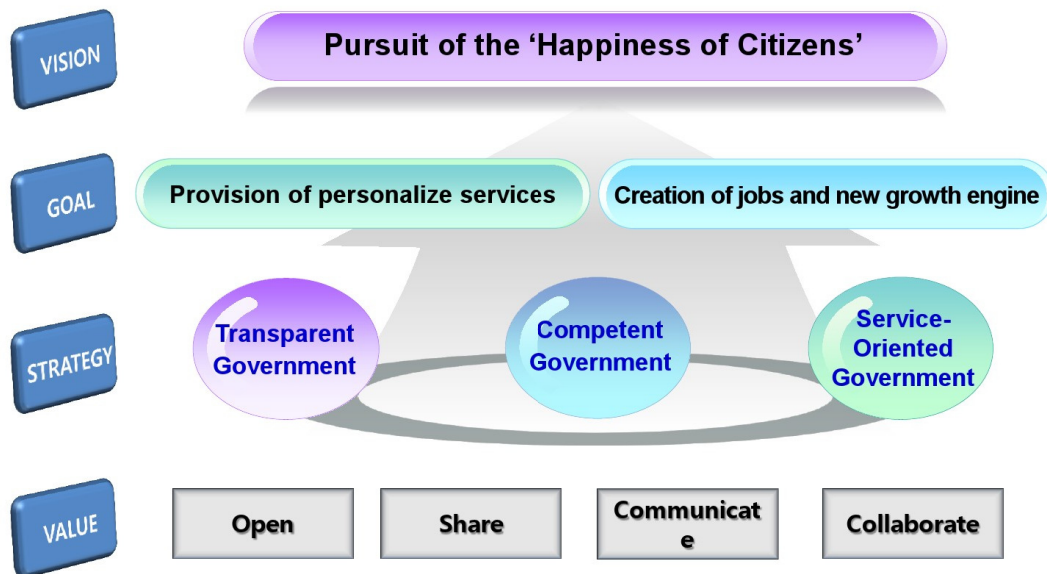
5

## Obama's Open Government Directive



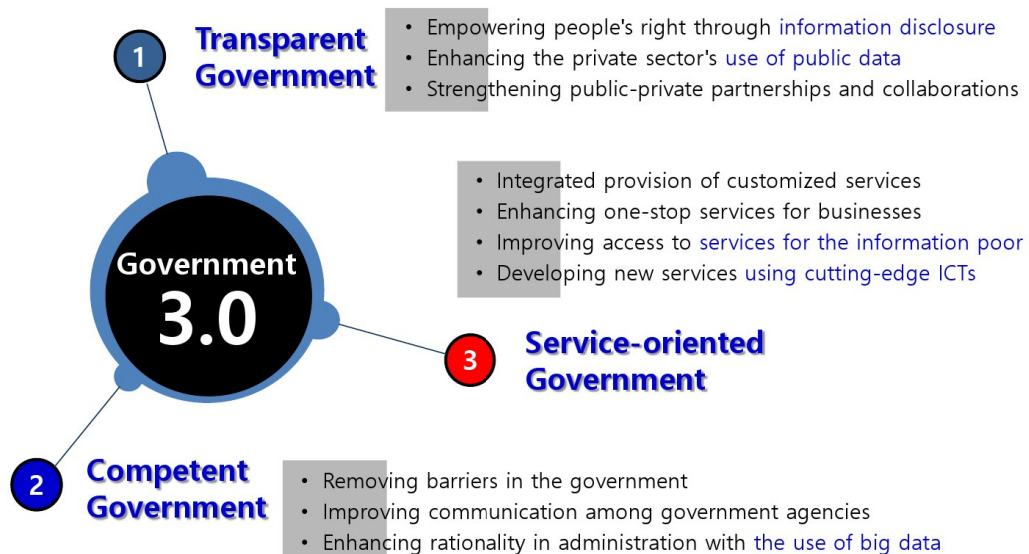
6

## Elements of Government 3.0



7

## Strategy of Government 3.0



8

# Open Data Strategy Council



9

# Open Data Portal (data.go.kr)



10



# Open Data Mediation Committee

**위원회 소개**  
공공데이터제공분쟁조정위원회는 2013년 10월 31일부터 공공데이터제공 및 활용에관한법률 제28조에 의거하여, 공공기관의 공공데이터 제공거부 및 제공중단에 관한 분쟁을 조정하기위하여 만들어진 기구입니다.

**공공데이터제공분쟁조정 제도안내**  
공공데이터제공 분쟁조정제도는 공공기관이 보유하고 있는 데이터의 제공거부 및 제공중단에 대하여 분쟁이 발생하였을 때, 전문성 있는 위원들로 하여금 다양하고 전문적인 분쟁의 세부적인 타협점을 모색 하도록 하여 국민의 공공데이터 이용 가능성을 높이고자 하는 취지에서 만들어진 제도입니다.

**분쟁조정 절차**  
공공데이터의 제공거부 및 제공중단을 받은 사람등 그 처분에 불복하여 80일 이내에 분쟁조정위원회에 분쟁조정을 신청할 수 있으며, 부득이한 사유가 있는 한 신청일로부터 30일 이내에 이를 심사하게 됩니다.

**분쟁조정 사례** (+ 14/92)  
[Q A] 공공데이터가 영업비밀에 해... 2014-03-03  
[Q A] 인용하는 참고문헌의 저자가... 2014-02-26  
[Q A] 공공데이터의 열람적 이용 .... 2014-02-25  
[Q A] 공공기관 발간 보고서가 공... 2014-02-25

**공공데이터 관련동향** (+ 15/97)  
"공공DB, 공개 넘어 활용성 높여야..." 2014-04-04  
[이슈분석]박정국 저작권정책관 "공공...2014-04-03  
공공데이터 품질관리 강화 -디지털타임...2014-03-28  
NIA 공공데이터 불만 해결 나선다... 2014-03-27

**분쟁조정 신청안내** **분쟁조정 신청하기**

11

# Korea Big Data Center

**공지사항** | 빅데이터 뉴스 (+ 더보기 +)  
· 빅데이터 분석활용센터 활동 교육 안내 2014-03-25  
· [공지] 데이터 기반 미래전략 컨설팅... 2014-03-17  
· [공지] 빅데이터 커리큘럼 참조모델 ... 2014-03-12  
· [공지] 2014년 빅데이터 활용 스마트... 2014-03-11  
· [공공신상 안내서 사전공개] 2014년 빅... 2014-03-03

**최신 데이터** (+ 더보기 +)  
방송, 통신 | 와이브로 상의 P2P 프로그램 실행 성...  
제공기관 : 서울대학교 | 유형 : TXT | 크기 : 2.7M8  
방송, 통신 | KT COR/고객정보 가공한 형태의 summa...  
제공기관 : KT | 유형 : 엑셀 | 크기 : 1.17GB

**자료실** | 활용사례 (+ 더보기 +)  
· [동네 슈퍼 '빅데이터'로 활로 찾다 : KBS ... 2014-04-02  
· 창조경제 실현을 위한 2013 빅데이터 국내 사례집 2014-03-17  
· 빅데이터 커리큘럼 참조모델 1.0 2014-03-14  
· 빅데이터 역량강화단모형(Big-CAT) 2014-03-14  
· 빅데이터 분석활용센터 이용설명회 자료 2014-03-03

**데이터 제공건수**  
17,023 건  
데이터 분류 건수  
의료, 보건 15102  
방송, 통신 343  
IT서비스 891  
산업, 경제 75  
일반행정 5  
기타 7  
월별 통계  
15,000  
10,000  
5,000  
9월 10월 11월 12월 1월

12

## Open Data Quality Management Center

공공정보 품질관리 지원센터  
Public Data Quality Management

공공정보 품질관리란 | 품질동향 | 품질진단 | 품질관리 교육 | 소통 참여 | 센터소개

상상 그 이상의 가치  
**Good Data!**  
모두가 믿고 사용할 수 있는 데이터 품질 확보  
공공정보 품질관리 지원센터가 함께합니다

품질관리개요 | 법·지침·매뉴얼  
품질진단안내 | 품질개선안내  
품질진단 컨설팅 | 전문가 상담

SNS 바로가기 | FACEBOOK | TWITTER

새소식 | 공지사항 | 보도자료

품질동향

커뮤니티

교육신청

품질자기진단

센터소개

13

## Korean Intellectual Property Office

특허청  
특허청 홈페이지

정보공개 | 열린마당 | 정보마당 | 정책마당 | 특허마당 | 특허청소개

정보공개

정보공개제도안내  
정보공개 신청/확인  
사전 정보공개  
정보목록  
정보공개방  
정책실명제  
신하기관 경영정보

공공데이터 개방

정보공개

공공데이터 개방

공공데이터제출신청

통합검색

검색

전체 | OPEN API | DATA (40)

데이터셋 (40)

검색결과 40 건

날짜 기준순 | 날짜 낮은순 | 정확도순 | 조회도순 | 페이지당 10건

【데이터셋】 【특허/실용신안】 KSIIC-PC 매핑 2014.04.04 조회수 : 32  
분류 : 과학기술 > 과학기술연구 | 기관 : 특허청 | 서비스유형 : 다운로드  
【특허/실용신안】 KSIIC-PC 매핑

14

## Seoul Open Data Portal (data.seoul.go.kr)

The screenshot shows the Seoul Open Data Portal interface. The header includes navigation links: Dataset, Open API, 카탈로그, 참여소통, 고객센터, 열린데이터광장?, and 전체메뉴. A search bar is at the top right. The main content area features a 'Data Visualization!' banner with a circular chart. Below the banner, there are sections for '데이터 시각화 시범 서비스 OPEN', 'Open API 활용 사례 공유', and '문화재'. A sidebar on the left lists various data categories. A search bar is located at the top right.

15

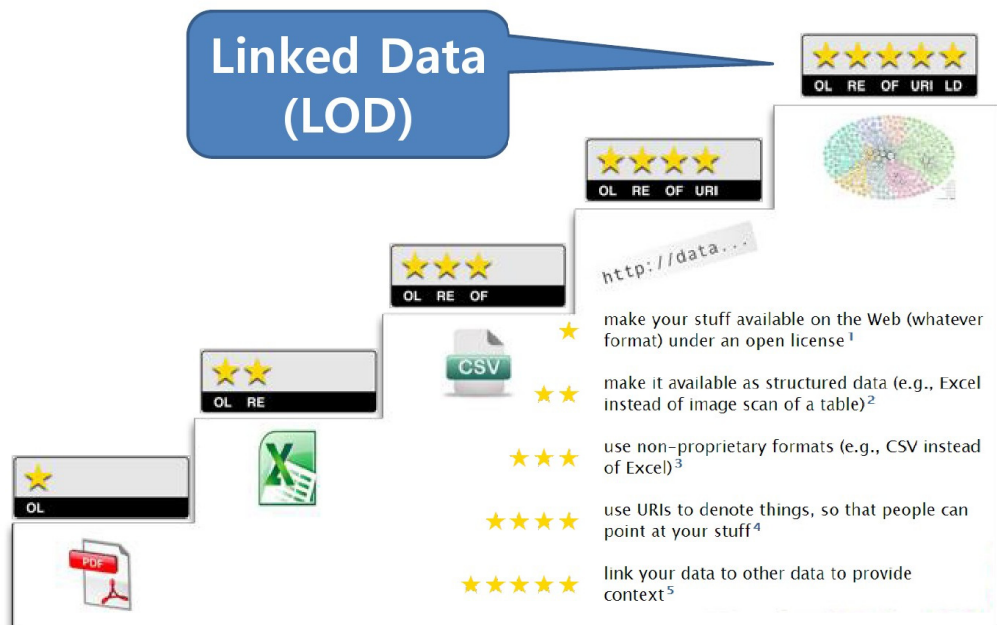
## GyeongGi Province Open Data Portal (data.gg.go.kr)

The screenshot shows the GyeongGi Province Open Data Portal interface. The header includes navigation links: HOME, 시행령 정보, 시행령 브라우징, and 지도기반 검색. A search bar is at the top right. The main content area features a 'GyeongGi Province Linked Open Data' banner with a circular diagram. Below the banner, there are sections for 'Sample', '유형별 선택', and '미리 보기'. A sidebar on the left lists various data categories. A search bar is located at the top right.

16



## 5-Star Open Data



17

## Linked Open Data for Sharing Spatial Big Data

18



Pre-Historic Era  
(12,000BC~3,000BC)

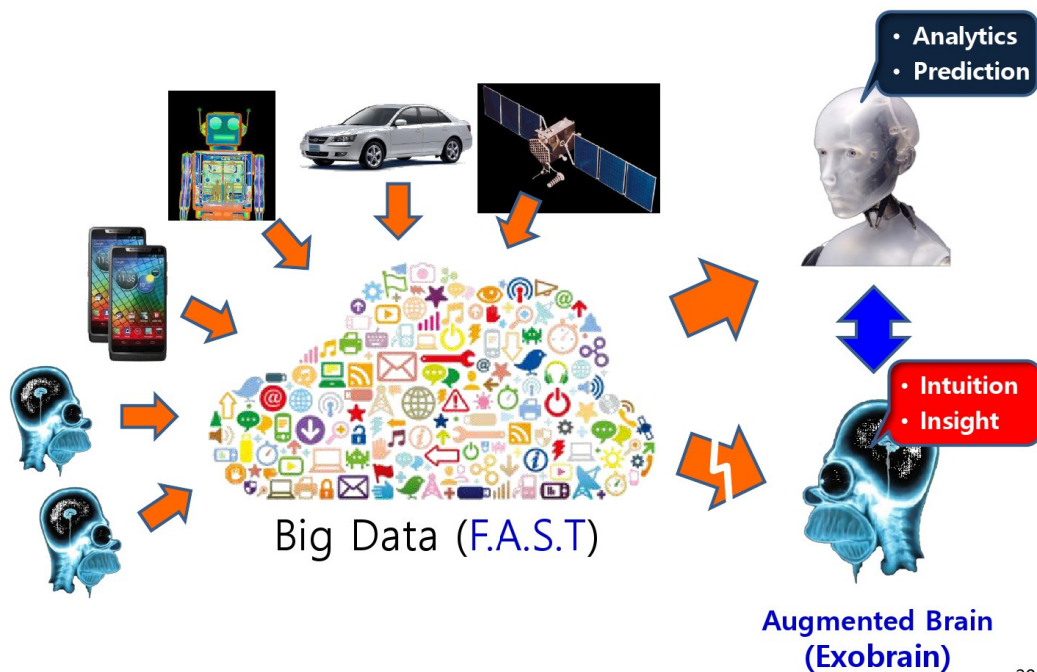


Historic Era  
( ~1,900AD)



19

Big Data Era ( 2000~ )



20

# What is Big Data?

## 3C ? 4C?

"Complex and large data sets that it becomes difficult to process using traditional technologies"

## F.A.C.T !!

(Fragment x Ambiguity x Context x Trustability)

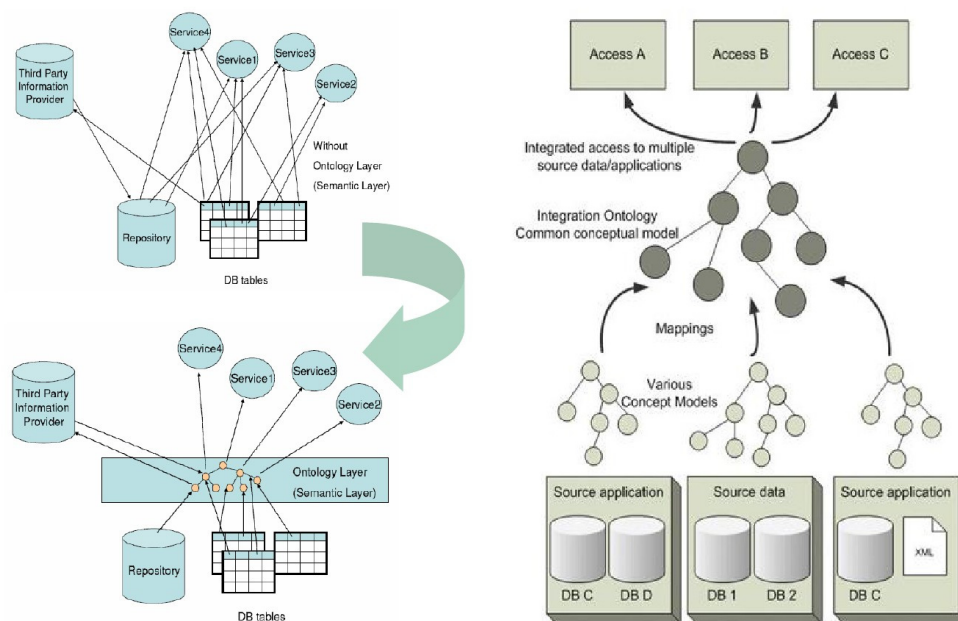
21

Data **Sharing** and Interoperability



22

## Data Sharing and Interoperability



23

## Common Language for Knowledge Sharing

Human	<b>Natural Language</b>	Human language written in letters: "The Earth orbits the sun in an ellipse"
	<b>Visual Language</b>	Visual expression of knowledge in picture, structure diagram, flow chart, and blueprint etc
	<b>Tagging</b>	Knowledge expressed in keywords, symbols and images related with objects
	<b>Symbolic Language</b>	Knowledge expressed in mathematical symbols : $x^2/a^2 + y^2/b^2 = 1$
	<b>Decision Tree</b>	Tree-shaped graph structure for complex decision making
Machine	<b>Rules Language</b>	Combined expression in condition with various rules of human knowledge
	<b>Database System</b>	Knowledge expression system composed of objects and relations in a table format
	<b>Logical Language</b>	Knowledge expression of logical symbols and arithmetic operations: Woman = Person $\cap$ Female
	<b>Frame Language</b>	Knowledge expression of values or pointers for other frames saved in slots
	<b>Semantic Network</b>	Knowledge expression of semantic relation between concepts in a graph structure
	<b>Statistical Knowledge</b>	Allows knowledge expression, machine learning technology combination based on probability and statistics

24

# Knowledge Representations

## Natural Language

"Employees working for a company are humans; the company and the employees are legal entities. The company is able to make a reservation for an employee's trip. The trip is available by plane or train that travels in cities within Korea or the U.S.. The companies and destinations for business trip are located in the cities. Saltlux reserved OZ510 with a round trip of Seoul and New York for Hong, Kildong."

## Rule Language

**(Rule)** If someone is flying, he must be on trip.

**(Rule)** If someone's trip is reserved in a company, he is an employee of the company.

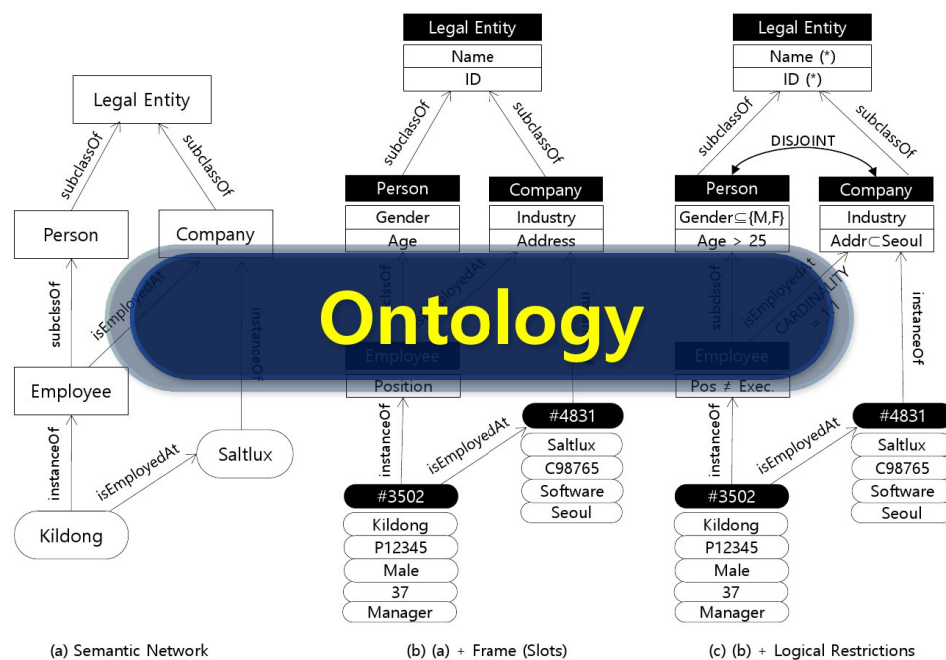
**(+ Rule)** For short trip in the same country, an employee should take a train.

**(Deduction)** Hong kil-dong whose flight is in reservation is an employee of Saltlux.

**(Deduction)** OZ510 is a flight for the U.S. and Korea.

25

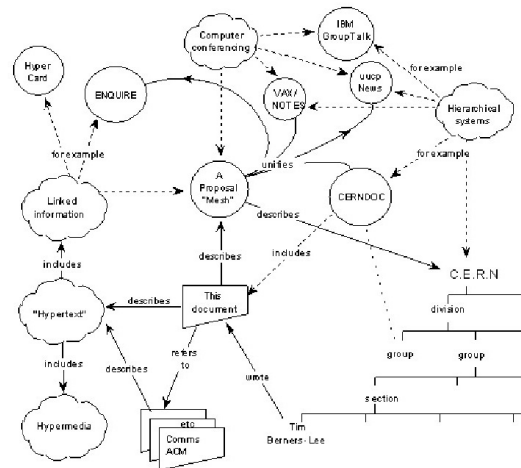
# Knowledge Representations



26

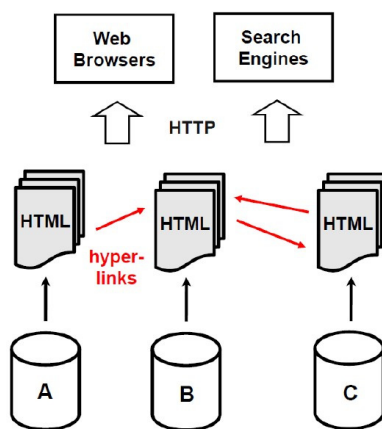


## World Wide Web and **Linked Data**

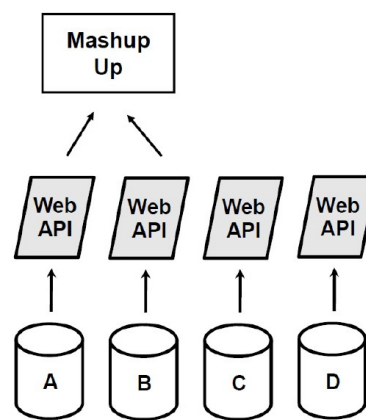


27

## The Web and Web 2.0



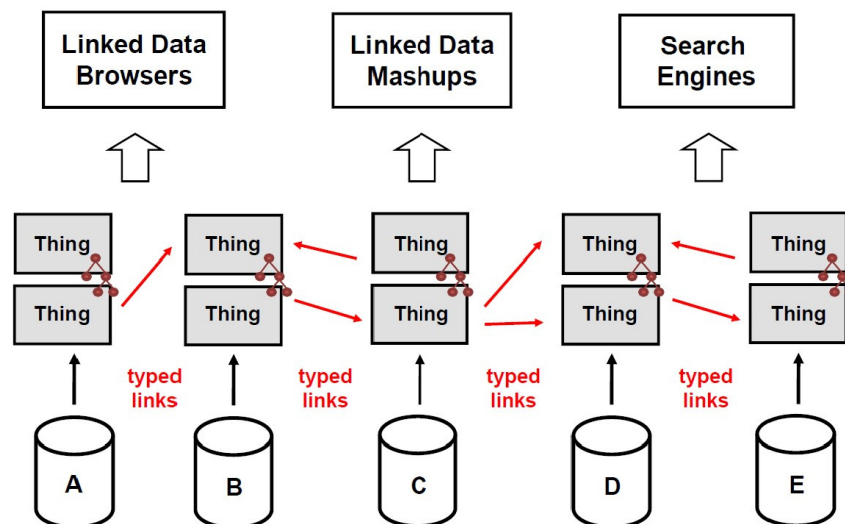
The Web



Web 2.0

28

## Linked Data Approach

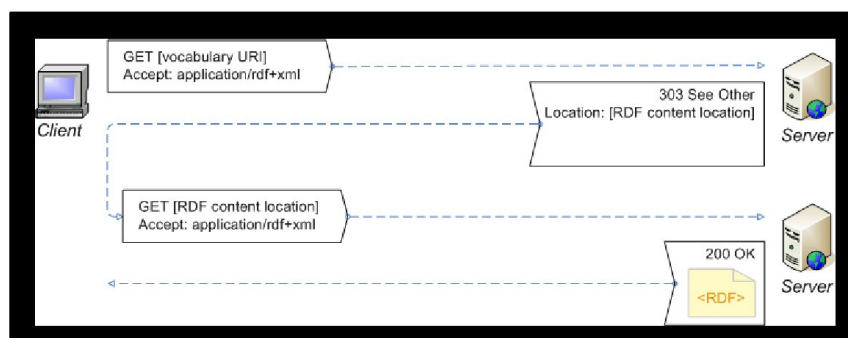


29

## Linked Data and **SPARQL** Endpoints

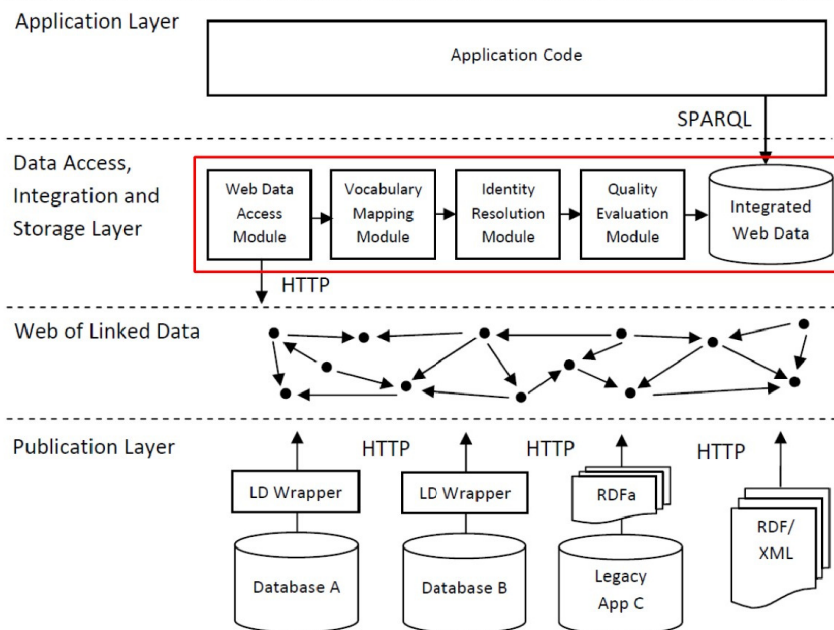
### REST protocol based semantic data querying on the Web

- Use HTTP URIs as names for things to look up those names.
- Use the standards (RDF(S), SPARQL)
- Include links to other URIs. to discover more things.



30

## Building Linked Data Applications



LOD2 Summer School, 2011

31

## 5 Key Features of Linked Data

### 1. Standard

W3C Standard based on Semantic Web, URI and HTTP protocol

### 2. Openness

Open data publishing, accessibility and sharing on the web

### 3. Flexibility

Easy data conversion and semantics working on RDF(S) and ontology

### 4. Interoperability

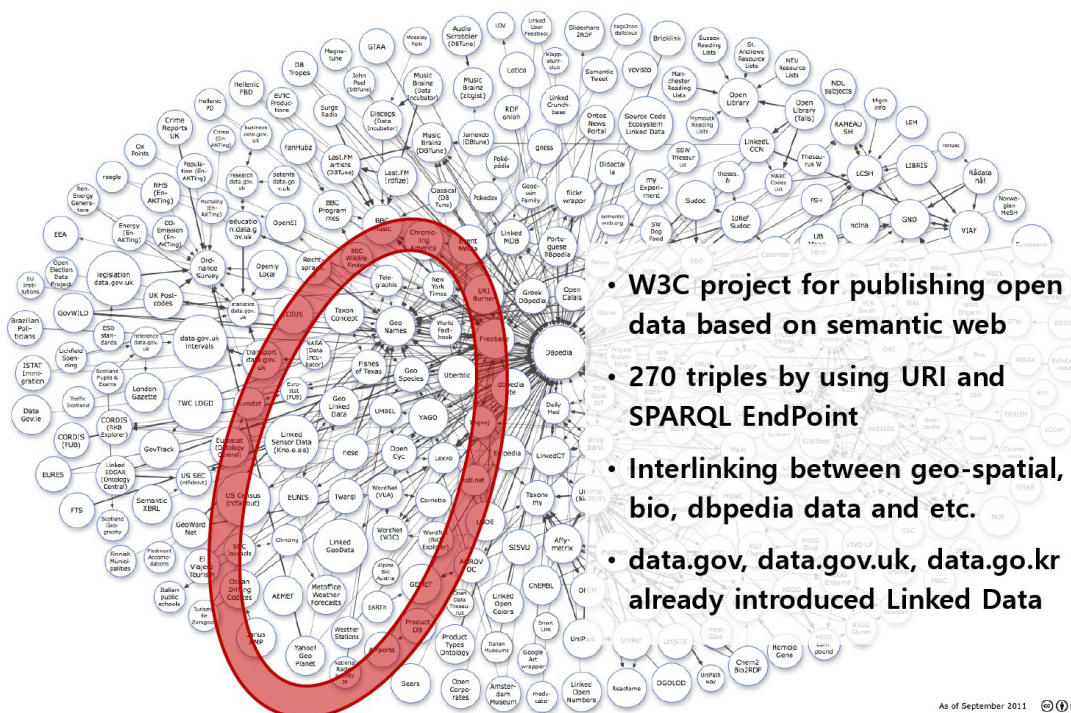
Data interoperability for heterogeneous data set by using data mash-up and powerful queries(SPARQL)

### 5. Machine readable

Machine can collect, read, store, query and infer distributed linked data

32

## LOD project (Linking Open Data)

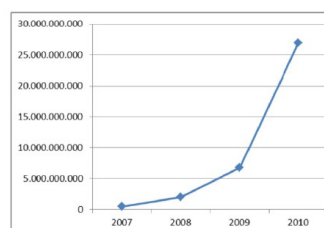


- W3C project for publishing open data based on semantic web
- 270 triples by using URI and SPARQL EndPoint
- Interlinking between geo-spatial, bio, dbpedia data and etc.
- data.gov, data.gov.uk, data.go.kr already introduced Linked Data

As of September 2011

## LOD project Statistics

Year	Datasets	Triples	Growth
2007	12	500.000.000	
2008	45	2.000.000.000	300%
2009	95	6.726.000.000	236%
2010	203	26.930.509.703	300%



Domain	Data Sets	Triples	Percent	RDF Links	Percent
Cross-domain	20	1,999,085,950	7.42	29,105,638	7.36
Geographic	16	5,904,980,833	21.93	16,589,086	4.19
Government	25	11,613,525,437	43.12	17,658,869	4.46
Media	26	2,453,898,811	9.11	50,374,304	12.74
Libraries	67	2,237,435,732	8.31	77,951,898	19.71
Life sciences	42	2,664,119,184	9.89	200,417,873	50.67
User Content	7	57,463,756	0.21	3,402,228	0.86
	<b>203</b>	<b>26,930,509,703</b>		<b>395,499,896</b>	



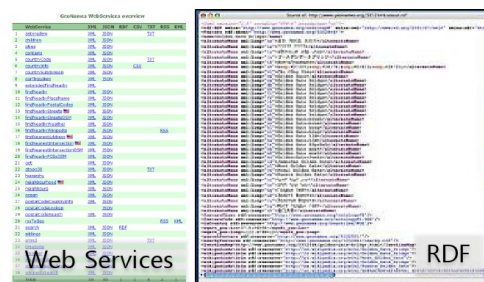
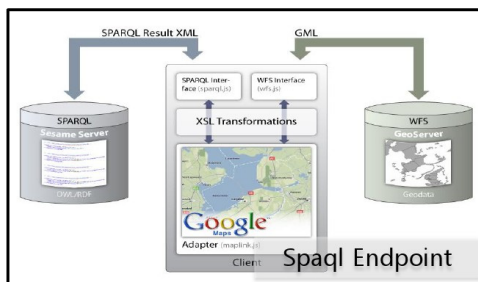
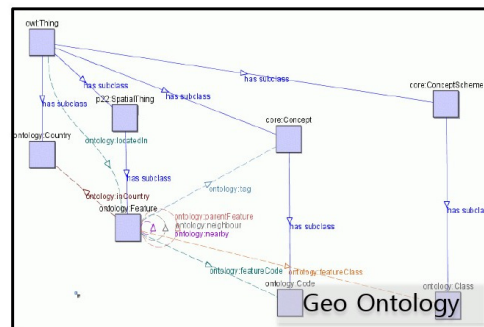
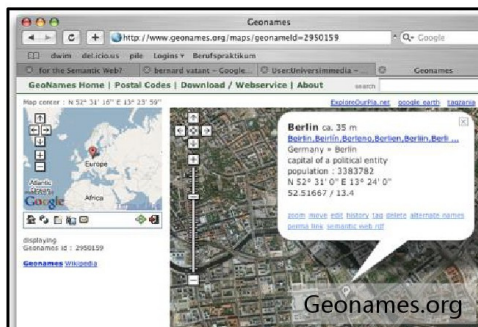
## Why **Geo-Spatial** Data for LOD?

It's a KEY for data interlinking and powerful applications



35

## GeoNames and GeoOntology



36



## OGC

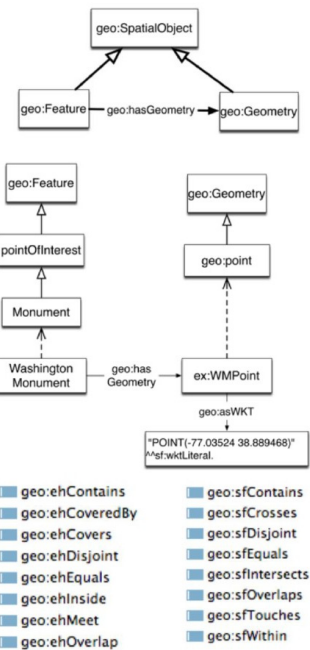
```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#" xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:geo="http://www.opengis.net/ont/geosparql#"
  xml:base="http://www.opengis.net/ont/geosparql">
```

GeoSPARQL 1.0 is an OGC Standard.  
Copyright (c) 2012 Open Geospatial Consortium.  
To obtain additional rights of use, visit <http://www.opengeospatial.org/legal/>.

Version: 1.0.1

```
<owl:Ontology rdf:about="">
  <dc:source rdf:resource="http://www.openis.net/doc/IS/geosparql/1.0/">
  <dc:source rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    OGC GeoSPARQL - A Geographic Query Language for RDF Data OGC 11-052-5
  </dc:source>
  <dc:date rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2011-04-30</dc:date>
  <owl:imports rdf:resource="http://www.openis.net/ont/gml/">
  <owl:import into="rdf:datatype"="http://www.w3.org/2001/XMLSchema#string">OGC GeoSPARQL
    1.0</owl:import into="rdf:datatype"="http://www.w3.org/2001/XMLSchema#string">
  <owl:imports rdf:resource="http://www.openis.net/ont/sf/">
  <dc:description rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    An RDF/XML vocabulary for representing spatial information
  </dc:description>
  <dc:source rdf:resource="http://www.openis.net/doc/full/ogc-geosparql/1.0/">
  <owl:imports rdf:resource="http://purl.org/dc/dolements/1.1/">
  <dc:source rdf:resource="http://www.openis.net/doc/IS/geosparql/1.0/">
  <dc:creator rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Open Geospatial
    Consortium</dc:creator>
  <owl:imports rdf:resource="http://www.w3.org/2004/02/skos/core/">
  <dc:source rdf:resource="http://www.openis.net/doc/full/ogc-geosparql/1.0/">
  </owl:Ontology>

  <rdf:type rdf:ID="vktLiteral">
  <dc:date rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2011-06-16</dc:date>
  <dc:comment xml:lang="en">
    A Well-known Text serialization of a geometry object.
  </dc:comment>
  <dc:description xml:lang="en">
    A Well-known Text serialization of a geometry object.
  </dc:description>
  <skos:definition xml:lang="en">
    A Well-known Text serialization of a geometry object.
  </skos:definition>
  <dc:isDefinedBy rdf:resource="">
  <skos:prefLabel xml:lang="en">Well-known Text Literal</skos:prefLabel>
  <dc:isDefinedBy rdf:resource="http://www.openis.net/spec/geosparql/1.0/">
  <dc:creator>OGC GeoSPARQL 1.0 Standard Working Group</dc:creator>
  <dc:label xml:lang="en">Well-known Text Literal</dc:label>
  <dc:contributor>Matthew Perry</dc:contributor>
  </rdf:Datatype>
```



[http://schemas.opengis.net/geosparql/1.0/geosparql\\_vocab\\_all.rdf](http://schemas.opengis.net/geosparql/1.0/geosparql_vocab_all.rdf)

37

# Ordnance Survey : OpenData

[illegible]

- **Open data for Geo-information**
- **TBL was involved in**
- **Supporting SPARQL EndPoint**
- **Supporting RDF/XML, Turtle, JSON**
- **Reference Ontologies**
  - Spatial Relations Ontology
  - WGS84 Geo Positioning
  - Gazetteer Ontology
  - FOAF

<http://www.ordnancesurvey.co.uk/oswebsite/opendata/>

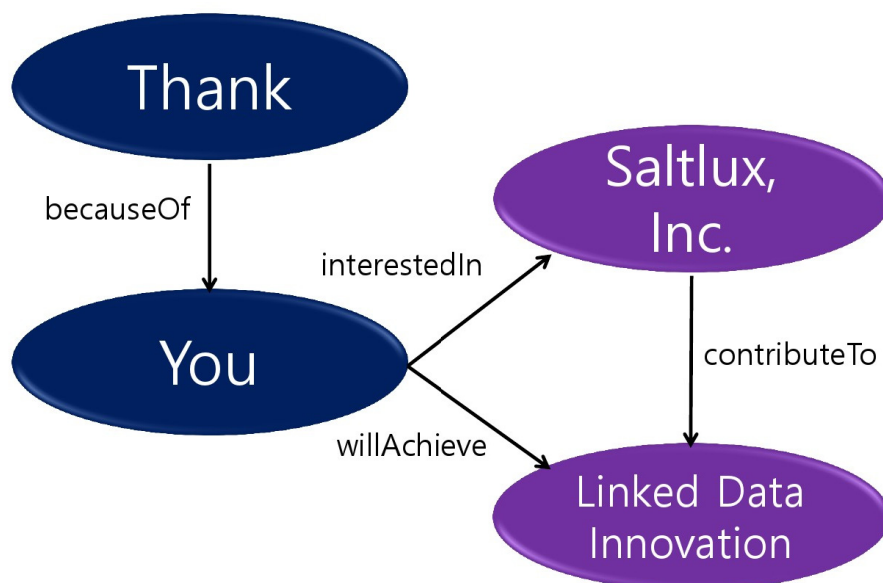
38

## Conclusion

"Computers are incredibly fast, accurate, and stupid.  
Human beings are incredibly slow, inaccurate, and brilliant.  
Together they are powerful beyond imagination."

- Albert Einstein -

- The era of human and machine collaboration.
- Healthy goose rather than big golden egg.



# Session 2

공간빅데이터 기술

Spatial Big Data Technologies

1. SpatialHadoop : A Map Reduce Framework for Spatial Big Data  
: Mohamed F. Mokbel  
(Professor, Univ. of Minnesota, USA)
2. Spatial Analytics Platform for Unstructured Big Data  
: Eui-Seon Jung(Director of Oracle Korea)
3. Cloud Technology for Effective Processing on Big Data  
: Byung-Gon Chun  
(Professor, Seoul National Univ, Korea)





## SpatialHadoop : A Map Reduce Framework for Spatial Big Data

Mohamed F. Mokbel (Professor, Univ. of Minnesota, USA)

### ABSTRACT

This talk is about SpatialHadoop; a full-fledged MapReduce framework with native support for spatial data. SpatialHadoop is a comprehensive extension to Hadoop that injects spatial data awareness in each Hadoop layer, namely, the language, storage, MapReduce, and operations layers. In the language layer, SpatialHadoop adds a simple and expressive high level language for spatial data types and operations. In the storage layer, SpatialHadoop adapts traditional spatial index structures, Grid, R-tree and R+-tree, to form a two-level spatial index. SpatialHadoop enriches the MapReduce layer by new components for efficient and scalable spatial data processing. In the operations layer, SpatialHadoop is already equipped with three basic operations, range query, kNN, and spatial join as case studies. Other spatial operations can also be added following a similar approach. The talk will also discuss various active projects for big spatial data that take advantage of SpatialHadoop.

### 1. Introduction

Since its release in 2007, Hadoop [1] was adopted as a solution for scalable processing of huge datasets in many applications, e.g., machine learning [10], graph processing [2], and behavioral simulations [14]. Hadoop employs MapReduce [7], a simplified programming paradigm for distributed processing, to build an efficient large-scale data processing framework. MapReduce abstracts distributed processing into map and reduce user defined functions. The map function maps each input record to a set of intermediate key-value pairs. The reduce function reduces intermediate similar-key records into a final result. Such abstraction simplifies the programming for developers, while the MapReduce framework handles parallelism, fault tolerance, and other low level issues. In the meantime, there is a recent explosion in the amounts of spatial data produced by various devices such as smart phones, satellites, and medical devices. For example, NASA satellite data archives exceeded 500 TB and is still growing [3]. Medical devices produce spatial images (X-rays) at a rate of 50 PB per year [4]. Such large-scale spatial data calls for taking advantage of the widely used Hadoop and MapReduce environments to support efficient spatial data querying and analysis. Unfortunately, Hadoop is ill-equipped to support spatial data as it deals with spatial data in the same way as non-spatial data.

In this talk, I will present SpatialHadoop; a full-fledged MapReduce framework with native support for spatial data. SpatialHadoop is built-in Hadoop as a comprehensive extension to Hadoop base code that pushes spatial constructs and spatial data awareness inside Hadoop core functionality. This results in allowing MapReduce programs and frameworks running on top of SpatialHadoop to make use of its embedded spatial functionality to achieve orders of magnitude better performance. SpatialHadoop to Hadoop will be the same as spatial database management systems(SDBMS) [13] to traditional DBMSs, where the original functionality of Hadoop and DBMS is still preserved with the addition of a native support of spatial data. As relational DBMS is ill equipped for spatial data, SDBMS has stepped forward to provide orders of magnitude performance improvement for spatial data processing through spatial operations and spatial index structure. SpatialHadoop follows a similar approach where new spatial data types, spatial index structures, and spatial operators are provided as built-in functionality in Hadoop while preserving traditional functionality of Hadoop.

SpatialHadoop is available for a download as a free open-source system at: <http://spatialhadoop.cs.umn.edu/>. So far, it has been downloaded more than 75,000 times since its first release on March 2013.

## 2. System Overview

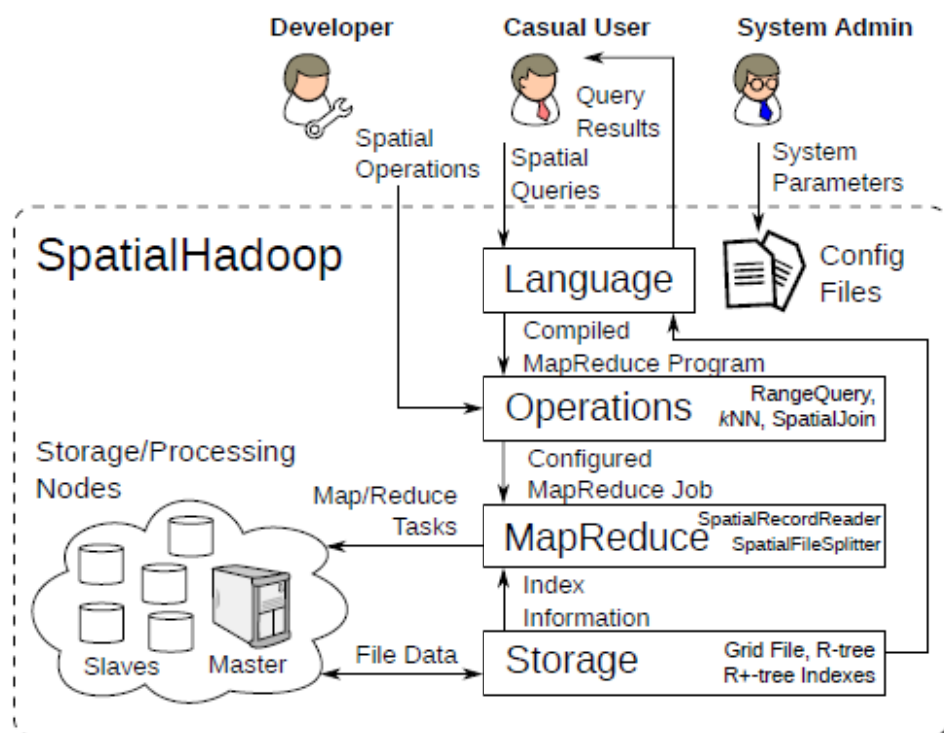


Figure 1: SpatialHadoop System Architecture

Figure 1 gives the high level architecture of SpatialHadoop. Similar to Hadoop, a SpatialHadoop cluster contains one master node that breaks a MapReduce job into smaller tasks, carried out by slave nodes. SpatialHadoop builds on this design by allowing the master node to plan a job with a minimal number of tasks and slave nodes to finish each task more efficiently. SpatialHadoop adopts a layered design of four main layers, namely, language, storage, MapReduce, and operations layers, while there are three types of users who interact with SpatialHadoop, namely, casual users, developers and administrators.

Figures 2a and 2b show how to express a spatial range query in Hadoop and SpatialHadoop, respectively. The query aims to find all points located within a rectangular area represented by two corner points  $\langle x1, y1 \rangle$  and  $\langle x2, y2 \rangle$ . The first query statement loads an input file of points, while the second statement selects records that overlap with the given range. This example distinguishes SpatialHadoop over Hadoop in two main aspects: (1) Performance: As Hadoop does not have any spatial indexes, it has to scan the whole dataset to answer the range query, which gives a very bad performance. In particular, it takes 200 seconds on a 20-node Hadoop cluster to process a workload of 60 GB. On the other side, our preliminary version of SpatialHadoop exploits its built-in spatial indexes to run the same query in about two seconds, which is two orders of magnitude improvement over Hadoop. (2) Readability: A Hadoop program, written in Pig Latin language [12], is less readable due to the lack of spatial data support in Hadoop high level languages. For example, our query in Hadoop uses the integer data type and numerical comparisons to express spatial operations. SpatialHadoop makes the program simpler and more expressive as it uses spatial data types (POINT and RECTANGLE) and spatial functions (IsOverlap). The readability issue becomes more serious with complex spatial data types, e.g., polygons.

```
Objects  = LOAD 'points' AS (id:int, x:int, y:int);
Result   = FILTER Objects BY x < xmax
          AND x > xmin AND
          y < ymax AND y > ymin;
```

(a) Range query in Hadoop

```
Objects  = LOAD 'points' AS (id:int, Location:POINT);
Result   = FILTER Objects BY
          IsOverlap (Location, RECTANGLE
                    (xmin, ymin, xmax, ymax));
```

(b) Range query in SpatialHadoop

Figure 2: RangeQuery in Hadoop vs. SpatialHadoop

### 3. Extensions to SpatialHadoop

#### 3.1 Pigeon

SpatialHadoop does not provide a completely new language. Instead, it provides, Pigeon [9], an extension to Pig Latin language [12] by adding spatial data types, functions, and operations that conform to the Open Geospatial Consortium (OGC) standard [5]. In particular, we add the following: (1) support for OGC-compliant spatial data types including, Point, LineString, and Polygon. Since Pig Latin does not allow defining new data types, Pigeon overrides the bytearray data type to define spatial data types. Conversion between bytearray and geometry is done automatically on the fly which makes it transparent to end users. (2) support for basic spatial functions, which are used to extract useful information from a single shape; e.g., Area calculates the area of a polygonal shape. (3) support for OGC standard spatial predicates, which return a Boolean value based on a test on the input polygon(s). For example, IsClosed tests if a linestring is closed while Touches checks if two geometries touch each other. (4) Spatial analysis functions that perform some spatial transformations on input objects such as calculating the Centroid or Intersection. These functions are usually used to perform a series of transformations on input records to produce final answer. (5) Spatial aggregate functions that take a set of spatial objects and return a single value which summarizes all input objects; e.g., the ConvexHull returns one polygon that represents the minimal convex polygon that contains all input objects.

#### 3.2 CG Hadoop

CG Hadoop [8] is a suite of computational geometry operations for MapReduce. It supports five fundamental computational geometry operations, namely, polygon union, skyline, convex hull, farthest pair, and closest pair, all implemented as MapReduce. In Hadoop, a computational geometry operation runs in three steps. (1) The partition step randomly partitions input records across nodes using the default Hadoop non-location-aware partitioner. (2) The local process step processes each partition independently and produces a partial answer stored as intermediate result. (3) In the global process step, the partial answers are collected in one machine which computes the final answer and writes it output. The drawback of Hadoop algorithms is that they need to scan the whole dataset.

In SpatialHadoop, we make two modifications to overcome this limitation. (1) we use the location-aware partitioner provided by SpatialHadoop in the partition step which groups nearby points in one partition. This allows us to run an extra pruning step before the local process step. In the pruning step, partitions that do not contribute to answer are early pruned without processing. Furthermore, the global process step also becomes more efficient as the size of its input (i.e., intermediate partial result) decreases.



## 4 Sample Projects

The core of Spatial Hadoop can be used to build scalable applications which deal with tons of spatial datasets. This section describes four key examples of systems that use SpatialHadoop as a powerful backend to handle spatial data processing.

### 4.1 MNTG: Minnesota Traffic Generator

MNTG [11], available at <http://mntg.cs.umn.edu/>, is a web-based traffic generator based on a real road network for the whole world. MNTG users can select an area on the map, specify a generation model and its parameters, then the system generates the traffic data on the backend and email back the user when the job is done. One challenge that MNTG faces is extracting the road network of the selected area before sending it to the generator. As the total size of the road network dataset is around 100 GB, a full scan would be tedious to do with every request. To overcome this problem, SpatialHadoop is used to construct an R+-tree index and use this index to speed up range queries on selected areas. In MNTG, we construct an index of around 10,000 partitions with an average partition size of 12 MB. This was experimentally found to be the best based on typical request sizes.

### 4.2 TAREEG : A Web Service for Extracting Spatial Data from OpenStreetMaps

TAREEG [6] is a web service for extracting spatial datasets from OpenStreetMap, available online at <http://tareeg.org/>. The interface is similar to MNTG, where a user selects an area on the map, chooses a dataset (e.g., road network), and submits an extraction request. On the back end, the server performs a range query on the selected data set and returns the extracted data in several formats including Google KML format and ESRI Shapefile. All the available datasets are extracted from OpenStreetMap using a MapReduce extractor that runs in SpatialHadoop. A Pigeon script is used to create points and connect them to form lines which form the shapes of the data (e.g., lakes and roads). After generating the datasets, SpatialHadoop is used to build R+-tree indexes, one per dataset, in order to speed up range queries. Total size of all datasets is around 400 GB.

### 4.3 SHAHED: A System for Spatio-temporal Analysis and Visualization of NASA Satellite Data

SHAHED is a tool for analyzing and exploring remote sensing data publicly available by NASA in a 500TB archive [3]. SHAHED provides a web interface where users navigate through the map and the system displays satellite data for the selected area. In addition, users can select an area and ask the system to display the change of temperature over a selected

time period as a video<sup>5</sup>). A user can also select an area and perform an analysis task on that area. For example, find anomalous patterns of vegetation in the selected area. This system uses SpatialHadoop to pre-compute the heat maps for available datasets and makes them available to the web browser for map navigation. The data mining module in the operations layer is also used to perform data analysis tasks issued by user.

#### 4.4 TAGHREED: A System for Querying, Analyzing, and Visualizing Twitter Data

Taghreed is a full-fledged system for efficient and scalable querying, analyzing, and visualizing geotagged microblogs, e.g., tweets. Taghreed supports arbitrary queries on a large number (Billions) of microblogs that go up to several months in the past. It consists of four main components: (1) Indexer, (2) query engine, (3) recovery manager, and (4) visualizer. Taghreed indexer uses SpatialHadoop to efficiently digest incoming microblogs. When the memory becomes full, a flushing policy manager transfers the memory contents to disk indexes which are managing Billions of microblogs for several months. On memory failure, the recovery manager restores the system status from replicated copies for the main-memory content. Taghreed query engine consists of two modules: a query optimizer and a query processor. Taghreed visualizer allows end users to issue a wide variety of spatio-temporal queries that exploit the index structures made by SpatialHadoop.

#### Speaker Bio

Mohamed F. Mokbel (Ph.D., Purdue University, USA, MS, B.Sc., Alexandria University, Egypt) is an associate professor at University of Minnesota as well as the founding Technical Director of the KACST GIS Technology Innovation Center, Umm Al-Qura University, Saudi Arabia. His current research interests focus on providing database and platform support for spatio-temporal data, location-based services 2.0, personalization, and recommender systems. His research work has been recognized by four best paper awards at IEEE MASS 2008, IEEE MDM 2009, SSTD 2011, and ACM MobiGIS Workshop 2012, and by the NSF CAREER award 2010. Mohamed is/was general co-chair of SSTD2011, program co-chair of ACM SIGSPATIAL GIS 2008–2010, and MDM2014, 2011. He has served in the editorial board of ACM Transactions on Spatial Algorithms and Systems, IEEE Data Engineering Bulletin, Distributed and Parallel Databases Journal, and Journal of Spatial Information Science. Mohamed has held various visiting positions at Microsoft Research, USA, Hong Kong Polytechnic University, and Umm Al-Qura University, Saudi Arabia. Mohamed is an ACM and IEEE member and a founding member of ACM SIGSPATIAL. He is currently serving as an elected chair of ACM SIGSPATIAL. For more information, please visit: <http://www.cs.umn.edu/mokbel>.

<sup>5</sup>) Please refer to an example at <http://youtu.be/hHrOSVAaak8>.

## References

- [1] <http://hadoop.apache.org/>.
- [2] <http://giraph.apache.org/>.
- [3] [https://lpdaac.usgs.gov/sites/default/files/public/modis/docs/MODIS\\_LP\\_QA\\_Tutorial-1.pdf](https://lpdaac.usgs.gov/sites/default/files/public/modis/docs/MODIS_LP_QA_Tutorial-1.pdf).
- [4] [http://www.eiroforum.org/activities/scientific\\_highlights/201209\\_XFEL/index.html](http://www.eiroforum.org/activities/scientific_highlights/201209_XFEL/index.html).
- [5] <http://www.opengeospatial.org/>.
- [6] Louai Alarabi, Ahmed Eldawy, Rami Alghamdi, and Mohamed F. Mokbel. TAREEQ: A MapReduce-Based Web Service for Extracting Spatial Data from OpenStreetMap. In SIGMOD, 2014.
- [7] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Communications of ACM*, 51, 2008.
- [8] Ahmed Eldawy, Yuan Li, Mohamed F. Mokbel, and Ravi Janardan. CG Hadoop: Computational Geometry in MapReduce. In SIGSPATIAL, 2013.
- [9] Ahmed Eldawy and Mohamed F. Mokbel. Pigeon: A Spatial MapReduce Language. In ICDE, 2014.
- [10] Amol Ghoting, Rajasekar Krishnamurthy, Edwin Pednault, Berthold Reinwald, Vikas Sindhwani, Shirish Tatikonda, Yuanyuan Tian, and Shivakumar Vaithyanathan. SystemML: Declarative Machine Learning on MapReduce.
- [11] Mohamed F. Mokbel, Louai Alarabi, Jie Bao, Ahmed Eldawy, Amr Magdy, Mohamed Sarwat, Ethan Waytas, and Steven Yackel. MNTG: An Extensible Web-based Traffic Generator. In SSTED, 2013.
- [12] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. Pig Latin: A Not-so-foreign Language for Data Processing. In SIGMOD, 2008.
- [13] Shashi Shekhar and Sanjay Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.
- [14] Guozhang Wang, Marcos Salles, Benjamin Sowell, Xun Wang, Tuan Cao, Alan Demers, Johannes Gehrke, and Walker White. Behavioral Simulations in MapReduce. PVLDB, 2010.

### **Abstract**

Many companies corporations thought that Big Data might be not valuable to them. However, they tried to utilize Big Data because they figured out that some other companies utilized Big Data and it was valuable. Nevertheless, most of companies underwent trial and errors because they had wrong prejudice that Big Data have to be based on Hadoop environments. Therefore, the most important thing to achieve their business goal is understanding variety of business requisites, and then applying matched solutions. When it comes to Spatial & Graph data, it is the same. By utilizing solutions that contain technique elements which satisfy business goal, companies can produce meaningful output. It was not easy to analyze Big Data, or Spatial & Graph data because of the limit of performance and meager development environment. In this sense, Oracle which tried to manage unstructured data added new feature for improvement of efficiency and upgraded Spatial & Graph analysis function to Oracle 12c. Plus, we need to consider Exadata as a Spatial & Graph data analysis platform because Exadata processes Big Data fast by In-memory database function, parallel processing skill, and highly efficient compressing data skill. Because of this new analysis function, it is possible to improve speed of join, or touch about 50 ~ 100 faster, and make application faster and easier to let in-database process handles massive data by database call. In addition, parallel raster operation will provide the performance improvement and simplified development environment. Also, by using added Virtual mosaicing, we are able to analyze the different format of images from another table with spatial query. Upgraded Spatial & graph feature are able to make NDM graph do real world feature modeling. Furthermore, we can get convenience that we can view existing relational table data and graph data through SPARQL because of RDF Semantic graph. I am assure that if we use the environment that contains improved function, and high-performance parallel processing function of Exadata, this will provide new environment and turning point to all of analysts of Spatial & Graph analysis.



## 비정형 빅데이터 공간정보의 분석을 위한 기반기술 (Spatial Analytics Platform for Unstructured Big Data)

2014. 08. 26

Euseon Jung, Director  
Business Development.  
Exadata & Strategic Solutions  
Oracle Korea

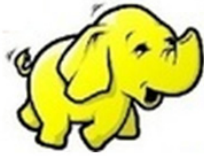
### Program Agenda

- 1 Spatial & Graph Data on Exadata
- 2 Oracle Spatial & Graph
- 3 Conclusion

ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

2



## Spatial & Graph Data on Exadata

- ✓ Oracle's Big Data Solutions
- ✓ Oracle DBMS & Unstructured data

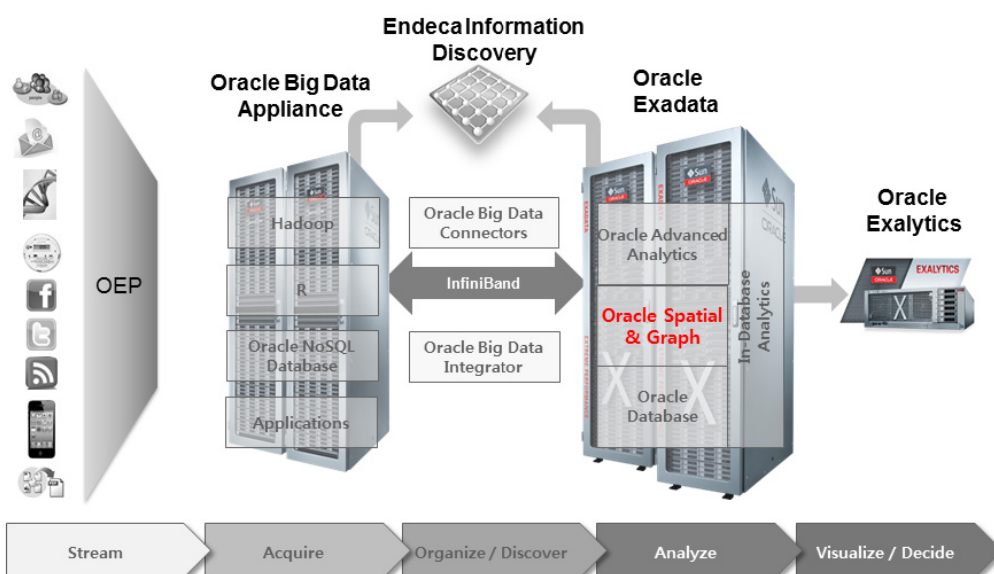


ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal/Restricted/Highly Restricted 3 3

## Oracle's Big Data Solution

Make Better Analysis Using Spatial & Graph Data



ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

4

## Oracle DBMS & Unstructured data

### ● Unstructured Data Support in Oracle Database

Oracle 8	Oracle 8i	Oracle 9i	Oracle Database 10g	Oracle Database 11g
VLDB	Text	XML DB	Network Graphs	SecureFile LOBs
LOBs	Spatial		RDF Graphs	DBFS
Extensibility	Image		Raster Imagery	DICOM Medical Imagery
	Audio		XQUERY	Oracle Spatial and Graph
	Video			BINARY XML

### ● Oracle Database Support for Unstructured Data

Optimized Storage

Specialized Data Types

Administration & Management

Indexing and Query

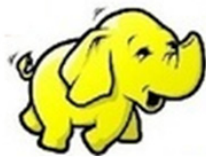
Powerful Analytics

**ORACLE®**  
DATABASE

**ORACLE®**

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

5



## Oracle Spatial & Graph



**ORACLE®**

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal/Restricted/Highly Restricted

6

## Advances with Oracle Database 12c

### New Spatial Features

Dramatic  
Performance



Simplified  
Application  
Development

ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

7

## Core Spatial Functions and Operations

### Performance

Oracle  
Database  
Locator

#### ANYINTERACT, INSIDE: 20-30x

DOES THIS FLOOD PLAIN HAVE ANY RELATIONSHIP WITH ANY OF THE PARCELS WE ARE INSURING?  
(any relationship can mean: is it inside the plain, does it touch the plain, etc.)

#### GEOM DISTANCE: 40X

HOW FAR IS HQ FROM THE BART STATION?

#### WITHIN DISTANCE: 10X

SHOW ME ALL THE RESTAURANTS WITHIN 5 MILES OF HQ.

#### VALIDATE GEOMETRY: 4X

This is a set of rules to check that what you are loading into Oracle's spatial types are "geometrically valid".

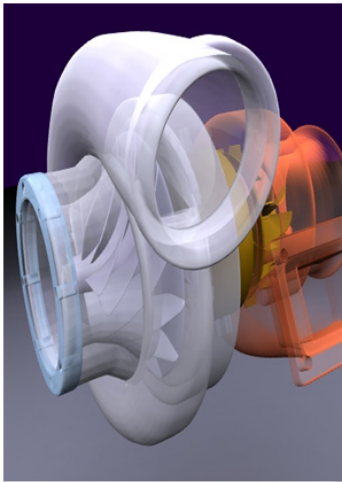
ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

8

## Vector Performance Acceleration

Oracle Spatial and Graph “Turbo-charger” feature



### OPTIMIZED METADATA QUERIES

- Kernel level caching
- Performance gains for DMLs and Spatial function calls
- Optimization especially noticeable in workflows with many fast running queries

ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

9

## Vector Performance Acceleration

“Turbo-charger” feature for spatial functions and operators

**Spatial & Graph  
option  
Performance  
Improvements**

### Join: 50-100x

JOIN is a database join except that the predicate is a spatial predicate, like WITHIN DISTANCE, INSIDE, etc. etc.)

### Touch: 50x

This is used in zoning and land management. DOES THIS PROPERTY TOUCH THE SHORELINE?

### Contains, Overlaps: 50x

Jurisdictional queries. Sales territory management. IS THIS HOUSE IN THIS SCHOOL DISTRICT? IF I PUT A NEW RETAIL STORE IN THIS LOCATION WILL ITS REACH OVERLAP WITH ANOTHER STORE'S?

### Complex masks: 50x

Masks are ways to combine multiple operators like "INSIDE + TOUCH". IS THIS PROPERTY INSIDE A FLOOD ZONE AND TOUCHING THE SHORELINE?

ORACLE®

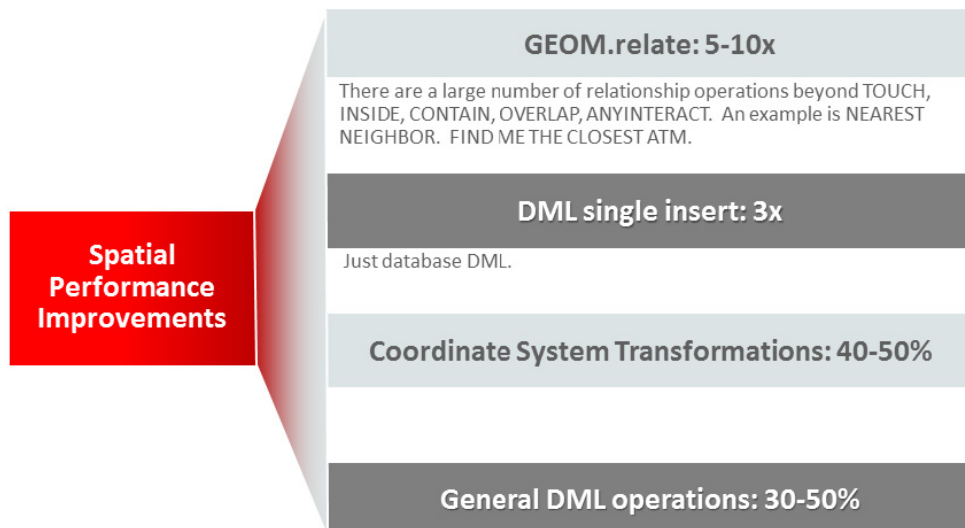
Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

10



## Vector Performance Acceleration

“Turbo-charger” feature for spatial functions and operations



ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

11

## New PointInPolygon Function

Fast Point in Polygon without Spatial index

- **SDO\_PointInPolygon** Function
  - Arg1: cursor that select a set of points
    - Very flexible as the data can come from a table, or result of another query
    - E.g., select \* from point\_data where c1 < 10 and c2 > 100 ...
  - Arg2: is any Polygon geometry
  - Returns all the points that are inside the polygon
- Useful when large number of points have to be classified based on a set of polygons
- Parallel enabled
- Can easily process 30K points per second in serial case

ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

12

## Parallel Raster Operations



- Many Raster functions can parallelize
- Serial operations perform up to 3x faster
- Scales to over 100x faster on highly parallel systems

ORACLE®

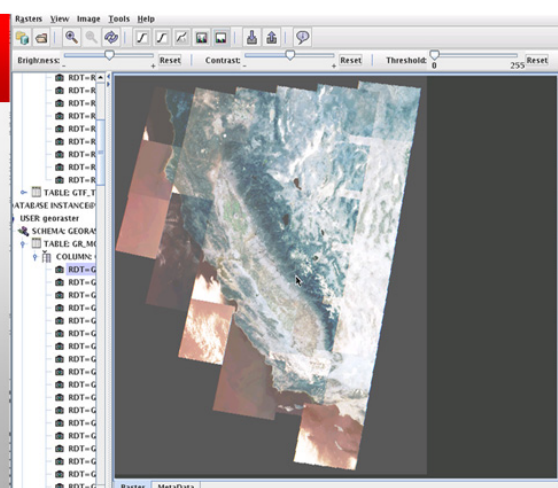
Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

13

## Virtual Mosaic and Image Processing

### In Database Processing

- Virtual Mosaic of collections of any georeferenced GeoRaster objects
- Advanced spatial queries and on-the-fly transformation and mosaics
- Raster Algebra operations to create new map products
- Image Processing: Masking, stretching, segmentation, rectification



Mosaic of Lands at Images

ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

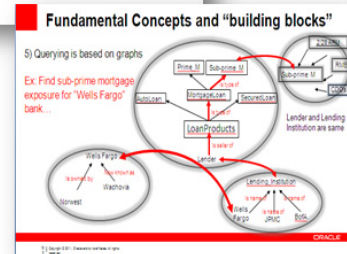
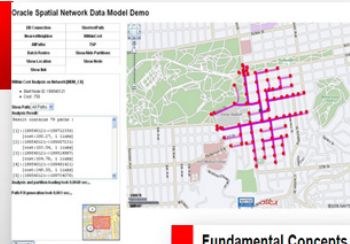
14

# Oracle Spatial and Graph

## Mature, Proven Graph Database Capabilities

### Graph Features

- Network Data Model graph
- W3C RDF Semantic graph



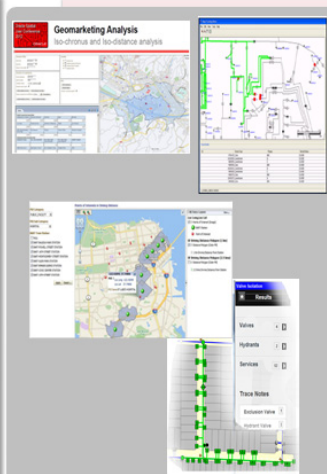
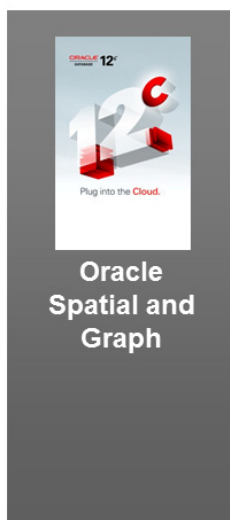
ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

15

## Network Data Model Graph

### Use Cases



- Transportation, Road and Multimodal Networks
- Drive Time Polygon Analysis
- Trade Area Management
- Service Delivery Optimization
- Water, Gas, Electric Utility, Network Applications

ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

16

## Oracle Spatial and Graph

### Network Data Model Graph

- A storage model to represent graphs and networks
  - Graph tables consist of links and nodes
  - Explicitly stores and maintains connectivity of the network graph
  - Attributes at link and node level
  - Logical or spatial graphs
  - Can logically partition the network graph
- Java API to perform Analysis in memory
  - Loads and retains only the partitions needed
  - Dynamic costs with real time input
  - Shortest path, within cost, nearest neighbors
  - Traveling salesman, spanning tree, ...
  - Multiple Cost Support in Path/Subpath Analysis

ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

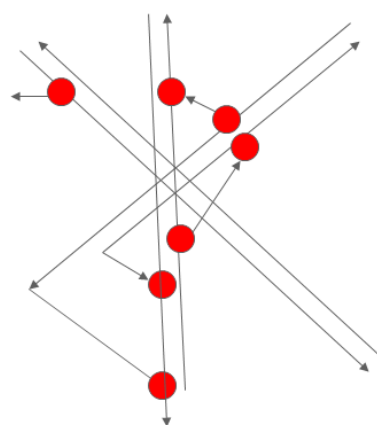
17

## Real World Feature Modeling in NDM Graph

Feature Representation



Network Representation



ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

18

## Network Data Model Graph

### Temporal Modeling/Analysis

- Traffic Patterns
  - Record historical travel
  - Based on time of day and day of the week
- NDM can use traffic patterns to compute shortest paths
- Support Nokia/HERE Traffic Patterns format out of the box

**Shortest Path Analysis**  
Left click for start point, right click for end point, or manually enter node ID, NWID, or geometry, or address.

Start: 15040037  
End: 15040135

**Network Constraints**  
(Hold ctrl key for multi-select or the select)  
Custom: NoHighwayConstraint  
Custom: ProhibitOneWayConstraint  
Oracle Spatial router mode: TrueWeightConstraint  
Oracle Spatial router mode: TrueLegalConstraint

**Prohibited Zone**

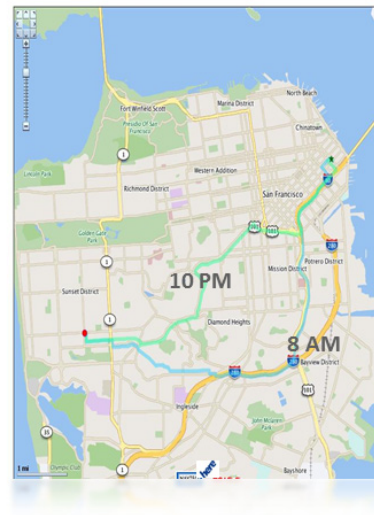
**Link Cost Calculators**  
Custom: TrafficLinkCostCalculator

**Keep Previous Results** ☒  
**Reverse Direction** ☐

**Include Traffic Data** ☒  
**Start Time** 10:00 PM  
**End Time** 8:00 AM

**Analysis Result:**  
(15040037-15040135)  
(15040135-15040037)  
Time to analyze the network: 0.487s.  
Time to compute geometries: 0.035s.

**Analysis Result:**  
(15040037-15040135)  
(15040135-15040037)  
Time to analyze the network: 0.436s.  
Time to compute geometries: 0.035s.



ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

19

## Network Data Model Graph

### Multi-Modal Routing

- Each mode (car, bus, rail, bike, etc) modeled as a separate network
- Single logical network represents all modes of transportation
- Transition nodes where networks meet
- NDM APIs can specify the modes
- Out of the box support for transit data published by transit authorities

**Analysis Result:**  
From: 575456205  
To: 575451525

**DriveWeb to:**  
"CONNECTICUT AV and WYOMING AV"  
(21 meters).

[1]  
Board Route 227 (busbound)  
At "CONNECTICUT AV and WYOMING AV"  
Dep. Time: 10:10:42

Get down at "NW CONNECTICUT AV and NW 20TH ST";

[2]  
Transfer to Route 80  
Board Route 80 (busbound)  
At "NW CONNECTICUT AV and NW 20TH ST"  
Dep. Time: 10:10:58

Get down at "NW 15TH ST and NW JACKSON PL";

[3]  
Transfer to Route 75  
Board Route 75 (busbound)  
At "NW 15TH and NW JACKSON PL"  
Dep. Time: 10:13:42

Get down at "SE INDEPENDENCE AV and SE 10TH ST";

[4]  
Transfer to Route 131  
Board Route 131 (busbound)  
At "Y CAROLX ST and SE 10TH ST"  
Dep. Time: 11:05:00

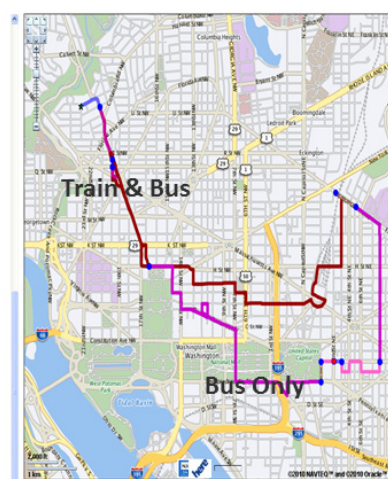
Get down at "Y CAROLX ST and SE 3RD ST"  
At 11:05:00

**DriveWeb from:**  
"Y CAROLX ST and SE 3RD ST"  
(0 meters) to destination.

**Trip Travel Time:** 51 minutes.

**Number of this Route=4**  
**Number of Train Routes=0**

**Time to analyze the network:** 0.579s.



ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

20

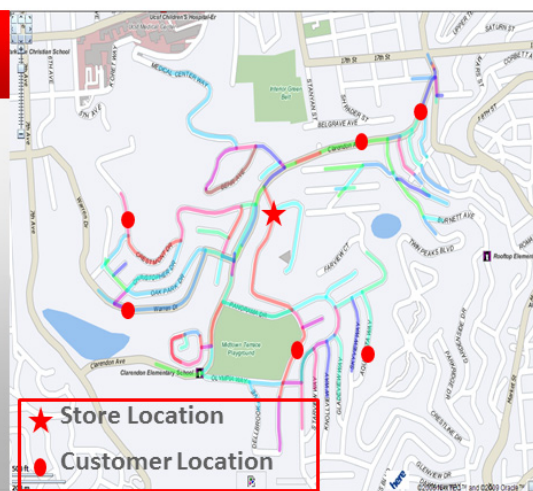


## Network Data Model Graph

Large Scale Drive Time/Distance Analysis

### Big Data Analysis

- Millions of customers, find closest store within a specified drive time
- Single database query to find closest store and drive time/distance for each customer
- Customers geocode as based on graph segment
- Network Buffer generates all possible paths



ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

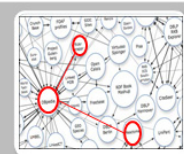
21

## RDF Semantic Graph

### Use Cases

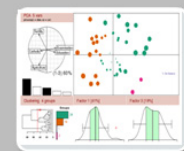
#### Linked Data & Public Clouds

- Unified content metadata model for public clouds
- Validate semantic and structural consistency



#### Text Mining & Entity Analytics

- Find related content & relations by navigating connected entities
- "Reason" across entities



#### Social Media Analysis

- Analyze content using integrated **metadata**
  - Blogs, wikis, video
  - Calendars, IM, voice



ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

22

## Oracle Spatial and Graph

## RDF Semantic Graph

## The Only RDF Database with:

- Support for both SPARQL and patented SQL access
- Works with OBIEE, Oracle BPM, Oracle Advanced Analytics
- Fine-grain Label-based Security

Conceptually, Semantic applications look at things as being represented as graphs, rather than tables

Type of Relationship	What you evaluate	What you compare	Opposite/Inverse Relationship
Lends to	Businesses and related parties	Businesses	Borrows from
Owls	Institutions and related parties	Institutions	Is owned by
Now known as	Corporate names and symbols	Corporate names	Previously known as
Operates in	Geographic hierarchy	Geographic name	No presence

In Oracle Database, we use Triples and Key relationships to represent nodes and links in the Graph.



ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved.

## Oracle Spatial and Graph

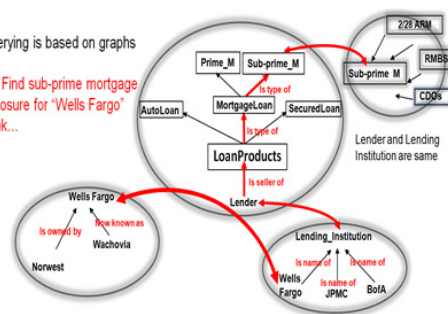
## RDF Semantic Graph

## Mature, complete RDF Database

- Supports all relevant W3C standards
- View relational data as RDF graph
- Scales with hardware— petabytes
- 60% data compression reduces storage and enhances performance

Querying is based on graphs

Ex: Find sub-prime mortgage exposure for "Wells Fargo" bank...



ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved.

24

## New functions in Oracle Database 12c

### RDF Semantic Graph

- RDF views on relational tables
  - RDF views can be created on a set of relational tables and/or views
  - SPARQL queries access data from both a relational and RDF store
  - Allows filtering of data in a relational store based upon graph analysis
  - Support RDF view creation using
    - Direct Mapping: simple and straightforward to use
    - R2RML Mapping: customizations allowed

R2RML : RDB to RDF Mapping Language

ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

25

## RDF Graph results with Oracle Business Intelligence SPARQL Gateway



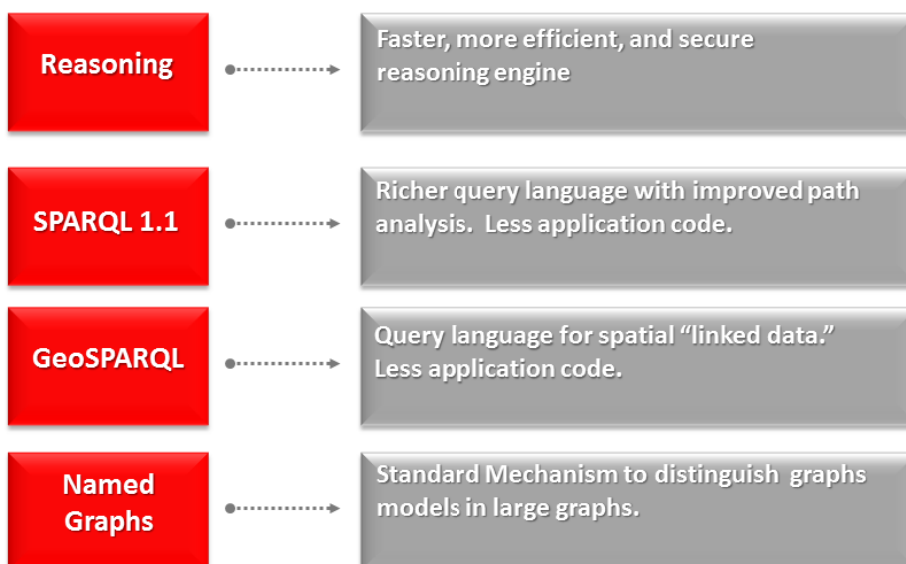
ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

26

## Performance and In-Database Analysis

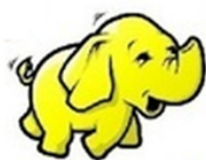
### RDF Semantic Graph



ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

27



## Conclusion



ORACLE®

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. | Oracle Confidential – Internal/Restricted/Highly Restricted

28

## Summary of New Spatial & Graph Features

- Vector Performance Acceleration
- High-performance point-in-polygon processing
- Parallel GeoRaster and Enhanced Raster Operations
- Network Data Model graph
  - ✓ Real World Feature modeling, multimodal Routing
  - ✓ Temporal Modeling and Analysis
  - ✓ Large Scale Drive Time/Distance Analysis
- RDF Semantic Graph
  - ✓ RDF views on relational tables
  - ✓ SPARQL 1.1, GeoSPARQL, SPARQL Gateway
  - ✓ Enhanced Reasoning and Security
  - ✓ Named Graphs

ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

29

## Oracle Exadata

Extreme Scalability for Millions of Spatial Objects

ORACLE  
EXADATA



- Millions of spatial objects evaluated in minutes
  - Point in polygon analysis
  - Polygon to polygon analysis
  - Deviation from route
  - Distance covered
- Millions of Spatial objects ingested in minutes
  - Weather readings
  - Traffic readings
  - Sensor readings

ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

30



## Oracle Exadata

### High Performance RDF Graph Workloads

ORACLE  
EXADATA



- RDF Semantic Graph is designed for the Exadata architecture
- 3x faster inferencing and querying
- Parallel load, inference and query
- Inferencing accelerated with Hybrid Columnar Compression
- Queries faster with OLTP index compression on B-tree indexes

ORACLE

Copyright © 2014 Oracle and/or its affiliates. All rights reserved. |

31

## REEF : Towards an Operating System for Big Data

Byung-Gon Chun (Professor, Seoul National Univ, Korea)

### Abstract

REEF (Retainable Evaluator Execution Framework) is a scale-out computing fabric that eases the development of Big Data applications on top of resource managers such as Apache YARN and Mesos. The resource management layer has emerged as a critical layer in the new scale-out data processing stack; resource managers assume the responsibility of multiplexing a cluster of shared-nothing machines across heterogeneous applications. They operate behind an interface for leasing containers – a slice of a machine’s resources – to computations in an elastic fashion. However, building data processing frameworks directly on this layer comes at a high cost: each framework must tackle the same challenges (e.g., fault-tolerance, task scheduling and coordination) and reimplement common mechanisms (e.g., caching, bulk transfers). REEF provides a reusable control and data plane for scheduling, coordinating, and executing task-level work on cluster resource managers. The REEF design enables sophisticated optimizations, such as container re-use and data caching, and facilitates workflows that span multiple frameworks. Examples include pipelining data between different operators in a relational system, retaining state across iterations in iterative or recursive data flow, and passing the result of a MapReduce job to a Machine Learning computation. REEF has been released as open-source under the Apache 2.0 License since January 2014.



# REEF: Towards an Operating System for Big Data

August 26, 2014

Byung-Gon Chun

Cloud and Mobile Systems Lab  
Computer Science and Engineering Department  
Seoul National University

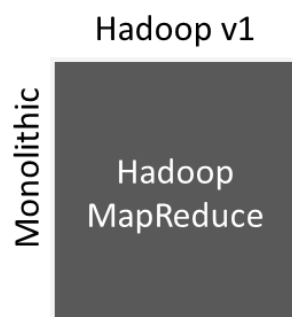
## REEF History

- 2012: Seeded at Microsoft
- Jan. 2014: Open-sourced under the Apache License 2
- Aug. 2014: Open-source Apache Incubator project (3<sup>rd</sup> Apache incubation that is driven by Korean committers)

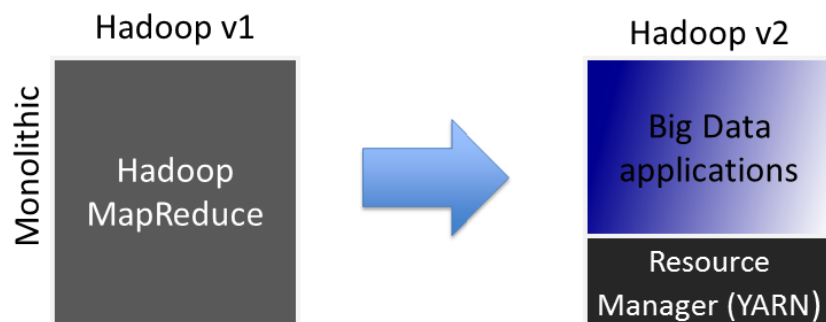
## REEF Team

- Seoul National University
- Microsoft
- UCLA
- SK Telecom
- UC Berkeley
- University of Washington
- Purestorage

## From a Monolithic Big Data Processing System

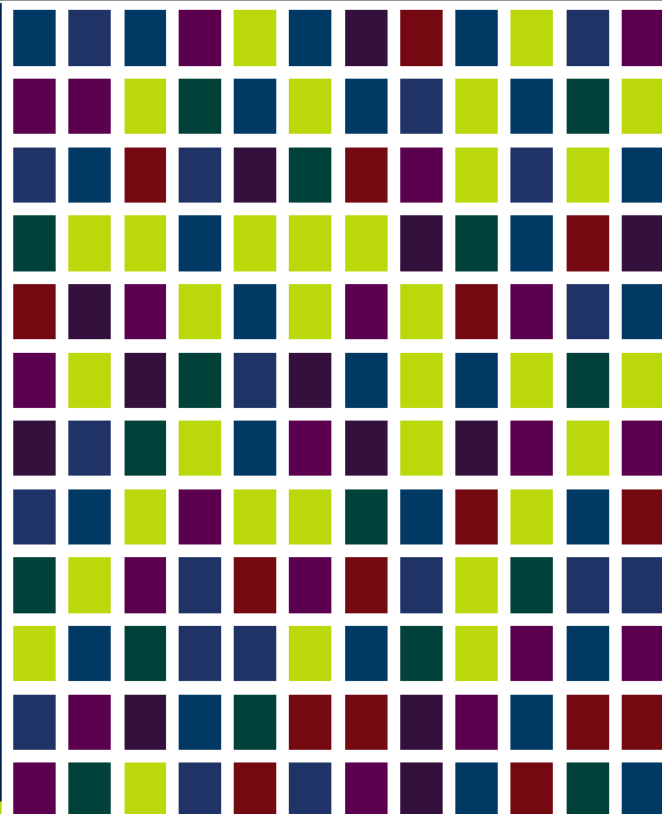


## A Recent Step Towards Refactoring Big Data Processing Systems



- Split up resource management and application job scheduling
- Split up resource management and computation models

Resource  
Managers







## Resource Managers

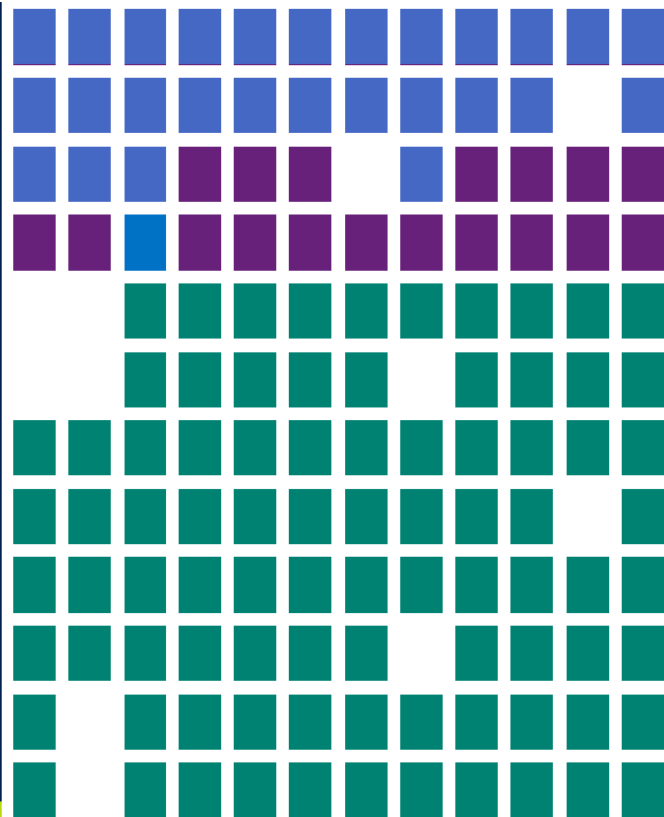
True multi-tenancy...

Many **workloads**: Streaming, Batch, Interactive, ...

Many **users**: Production Jobs, Ad-Hoc Jobs, Experiments, ...

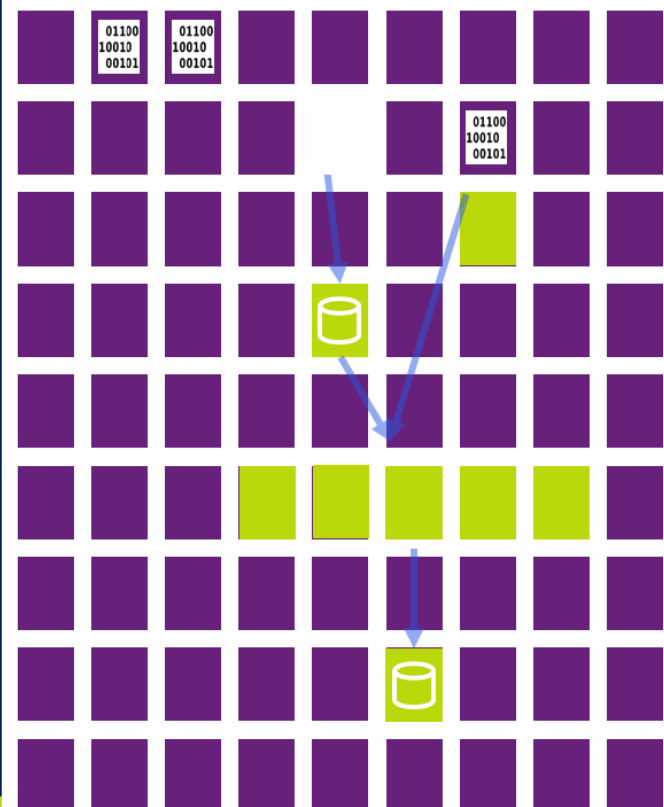
...but, only for sophisticated apps

Fault tolerance  
Pre-emption  
Elasticity



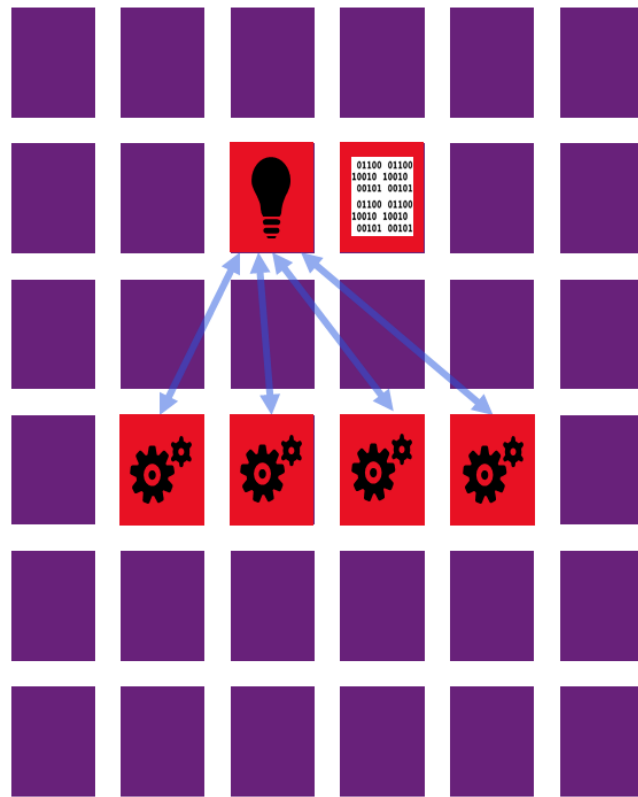
## Example 1: SQL / MapReduce

Fault tolerance  
Elasticity



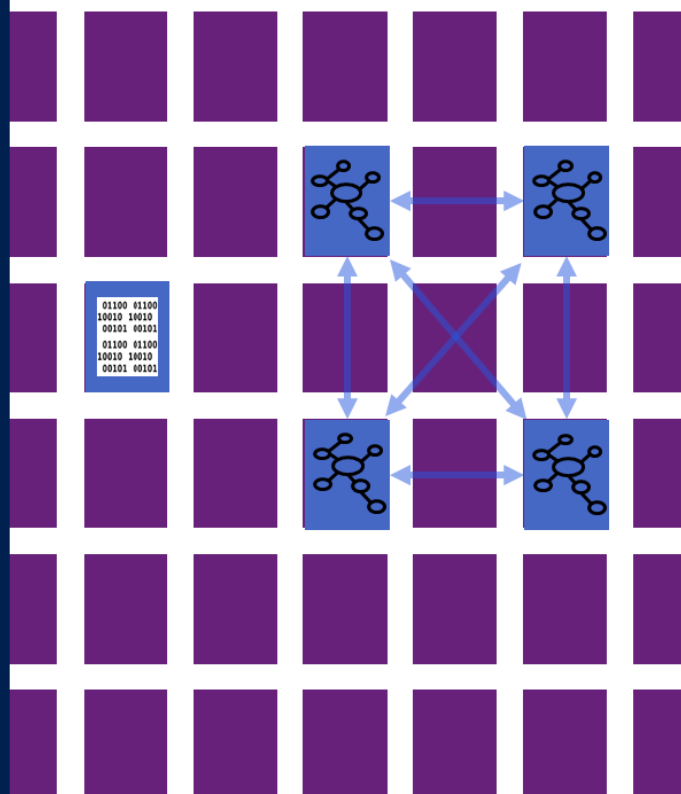
## Example 2: Machine learning

Fault tolerance  
Elasticity  
Iterative computations



## Example 3: Graph processing

Fault tolerance  
Elasticity  
Iterative computations  
Low latency communication



## Silos

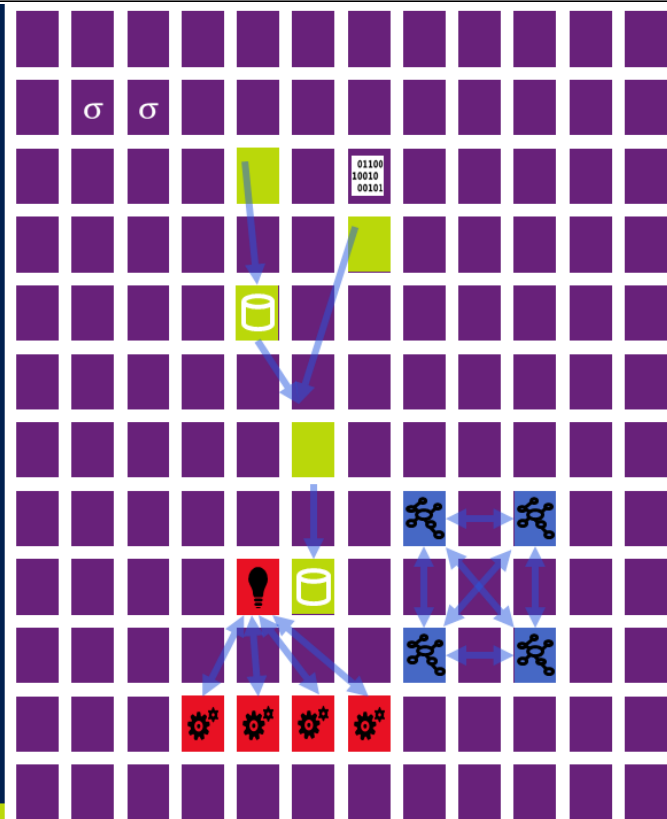
Silos are hard to build

Each duplicates the same mechanisms under the hood

In practice, silos form pipelines

In each step: Read from and write to HDFS  
Synchronize on complete data between steps

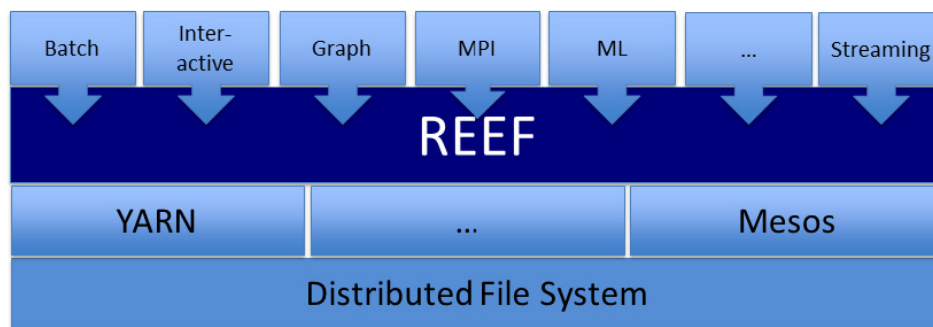
→ Slow



## REEF

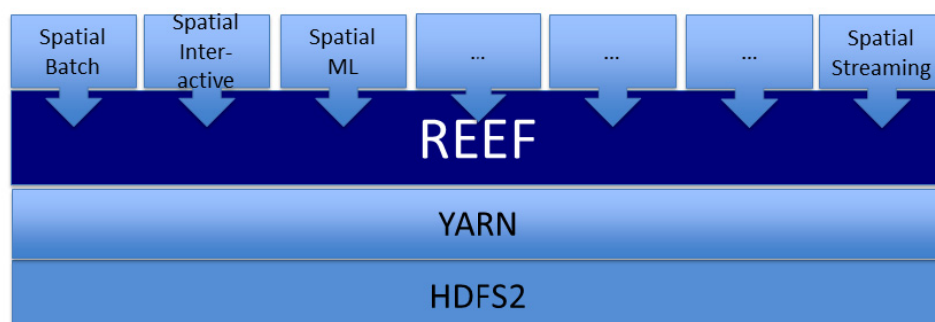
REEF (Retainable Evaluator Execution Framework) is a scale-out computing fabric that eases the development of Big Data applications on top of resource managers such as Apache YARN and Mesos.

## REEF Stack



- Common building blocks for heterogeneous Big Data applications - reusable control and data planes
- Virtualization of resource managers
- Container reuse, retained state across tasks from heterogeneous frameworks
- Simple configuration management and scalable event handling


## A Deployment Scenario








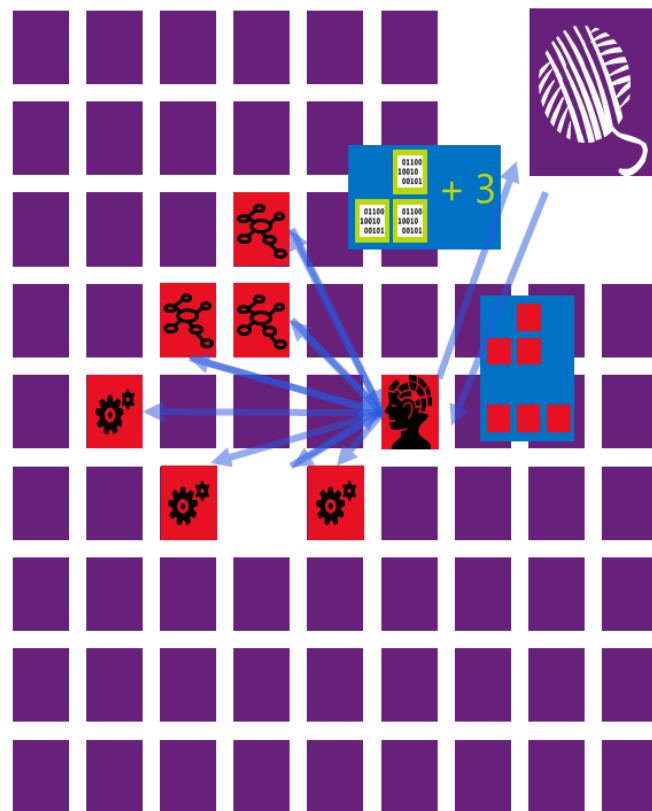


## REEF Control Flow

Yarn ( ) handles resource management (security, quotas, priorities)

Per-job REEF Drivers ( ) request resources, coordinate computations, and handle events: faults, preemption, etc...

REEF Evaluators ( ) hold hardware resources, allowing multiple REEF Tasks ( , , , ) etc.) to use the same cached state through REEF Contexts.

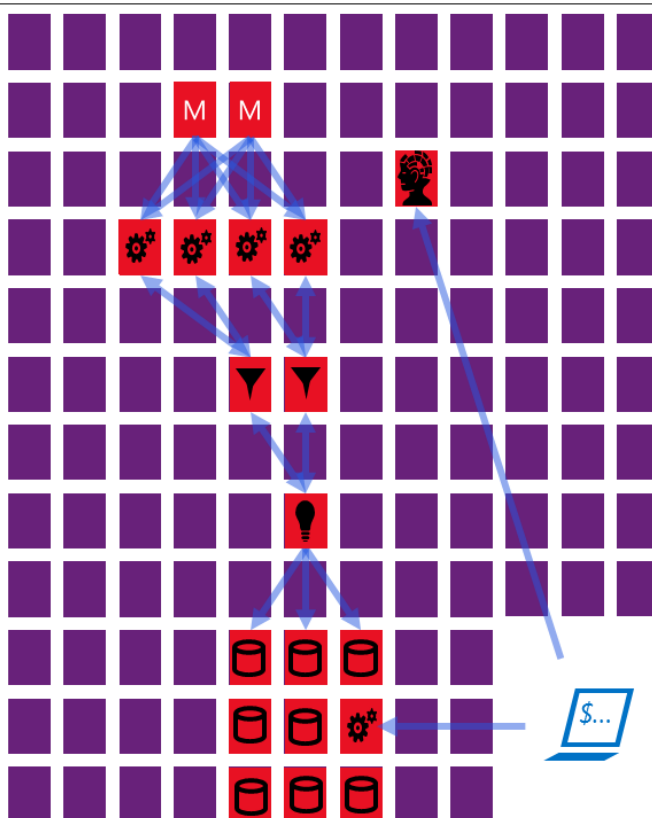


## Retaining Evaluators

Handover of pre-partitioned and parsed data between frameworks

Iterative computation

Interactive queries



# Control Flow Take-Away

## Easy to reason about

### Centralized control flow

Evaluator allocation & configuration  
Task configuration & submission

### Centralized error handling

Task exceptions are thrown at the Driver  
Evaluator failure is reported to the Driver

## Scalable

### Event-Based Programming

Driver fires requests as events to REEF  
REEF fires events to the Driver

### Mostly stateless design

REEF maintains minimal state  
Majority of the state keeping (e.g. work queues) is maintained by the Driver.

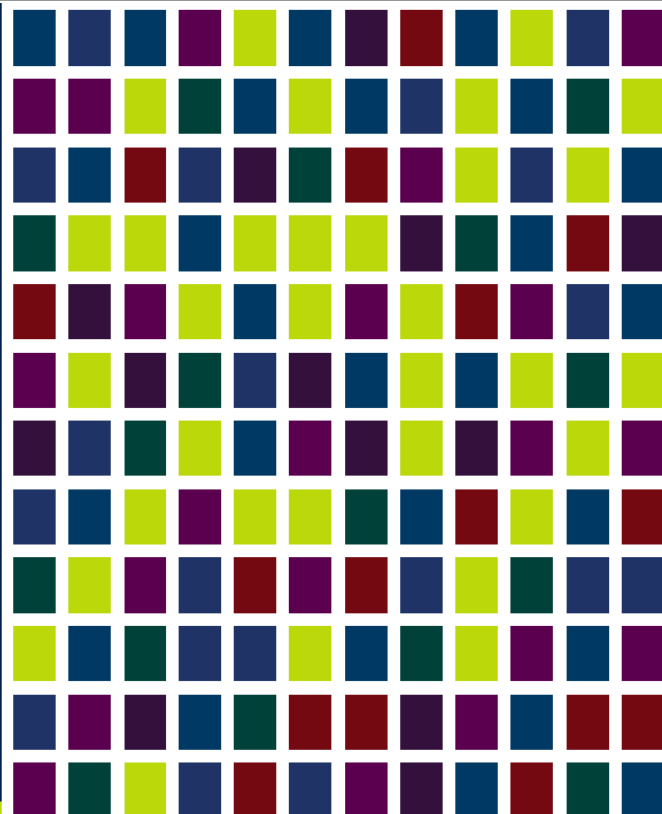
## Wake: Events + I/O

Event based programming and remoting

API: A static subset of Rx  
→ static checking of event flows

Thread management

Multiplexing network I/O  
Message ordering



Tang

## Configuration is hard

Errors often show up at runtime only  
State of receiving process is unknown to the configuring process

## Our approach

- Configuration as Dependency Injection
  - Configuration here is pure data
  - Early static and dynamic checks

## Design guideline

## Immutable & Commutative

```

Error:
  container=15023340523847-
02.stdout"
  YarnEvaluator
NullPointerException at:
Evaluator
  evaluate():1234
  ShellTask.helper():546
Error:
  ShellTask.onNext():789
  Unrequired parameter "command"
Missing required parameter
"cmd" Got ShellTask

```

# #!

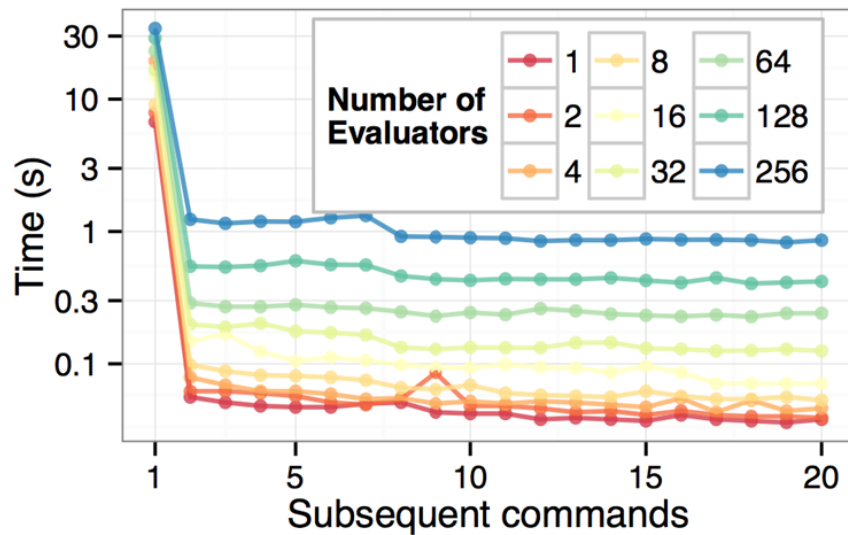
# Data Plane Libraries

## Storage: Map, Spool, ...

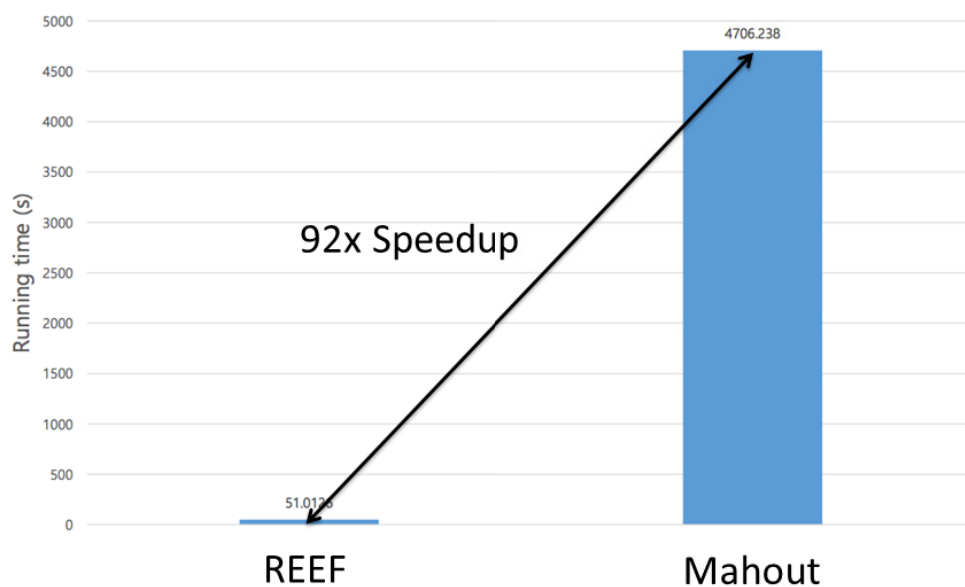
Network: Identity-based communication,  
Group communication, ...

## State management: Checkpoint

## Interactive Distributed Shell (Evaluator Reuse)

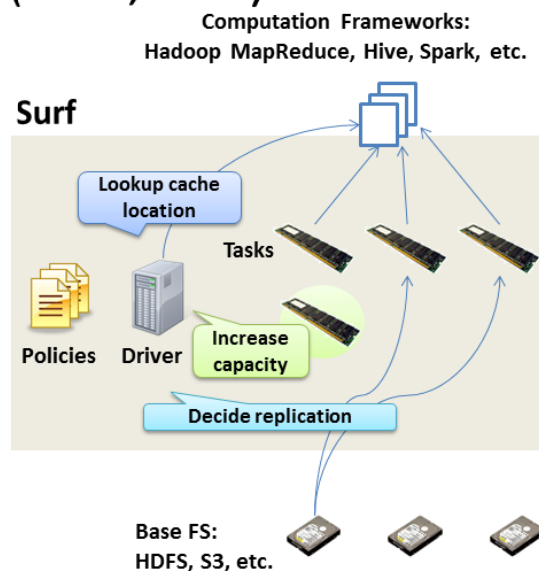


## k-means (Evaluator Reuse and Retained State)



## Surf: In-Memory Store for Big Data Analytics (SNU, SKT)

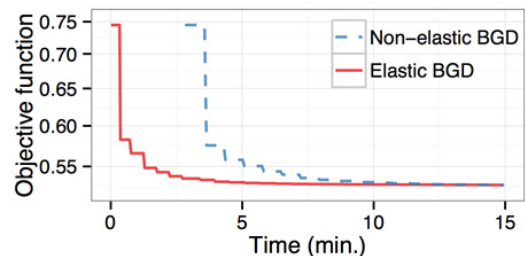
- In-memory distributed caching tier on REEF
  - Flexible configuration of policies
  - Elasticity
  - Easy read access across frameworks and base file systems



25

## Elastic Machine Learning (MS, SNU)

- Elastic group communication in REEF
  - Elastically add/remove nodes
  - Fault-awareness
- A new breed of elastic Machine Learning algorithms





## Other Developments

- Data movement abstraction
- Graph processing
- Stream processing
- New applications on REEF

## THANK YOU!

Contact

Byung-Gon Chun    [bgchun@snu.ac.kr](mailto:bgchun@snu.ac.kr)

Apache incubator project announcement  
<http://cmslab.snu.ac.kr/2014/08/15/reef-is-an-apache-incubator-project/>



# Session 3

공간빅데이터 활용사례

Spatial Big Data Applications

1. Big Data for Future Energy and Urban Infrastructures—Challenges and Opportunities  
: Budhendra Bhaduri  
(Director of Geographic Information Technologies Research Center, Oak Ridge National Lab, USA)
2. Applications of Micro Geo Data for Urban Monitoring  
: Yuki Akiyama  
(Research Fellow, Univ. of Tokyo, Japan)
3. How to use Big Data in LH  
: Yeon-Gurl Cho  
(Vice Director of Spatial Information Division, Korea Land & Housing Corporation)



## 1

## Big Data for Future Energy and Urban Infrastructures - Challenges and Opportunities

Budhendra Bhaduri

(Director of Geographic Information Technologies Research Center, Oak Ridge National Lab, USA)

### Abstract

In this rapidly urbanizing world, unprecedented rate of population growth is not only mirrored by increasing demand of energy, food, water, and other natural resources, but has detrimental impacts on environmental and human security. Much of our scientific and technological focus has been to ensure sustainable future with healthy people living on a healthy planet where energy, environment, and mobility interests are simultaneously optimized. Ability to observe and measure through direct instrumentation of our environment and infrastructures from buildings to planet scale, coupled with explosion of data from citizen sensors brings much promise for capturing the social/behavioral dimension and provides a unique opportunity to manage and increase efficiencies of existing built environments as well as design a more sustainable future. As a spatial research community, we must explore the intriguing developments in the world of Big Data and plausible ways citizens can all become part of the open data economy for advancing science and society.

### 1. Introduction

Under the current global concerns of energy scarcity and climate change, there is increasing realization that a transition from the current petroleum-dependent US society to one fueled by alternative clean energy sources, more efficient vehicles, and a more efficient transportation network is required for a prosperous sustainable future. In order to reduce fossil fuel dependence, environmental impacts, and congestion, a number of alternative energy supply, distribution, and end-use transportation systems, technologies and policies are presently being explored. These include conventional hybrid vehicles, Plug-in Hybrid Electric Vehicles (PHEVs), increased usage of biofuels, and IntelliDrive. Ideally, development and implementation of future strategies for alternative energy resources and technologies will assure a societal system in which energy, environment, and mobility interests are simultaneously optimized. Given the complex, intertwined nature of such system across geographic scales, assessing the effectiveness of possible planning strategies and discovering their unanticipated and unintended consequences require modeling and simulation utilizing finest resolution data, physical and social processes, and observing the emerging behavior of the system over large spatial and temporal scales.

## 2. Big spatial data challenge

A critical challenge has emerged from the explosion of observational and simulation data. The amount of visual and image data are increasing at the rate of terabytes to petabytes of data a day with the progress in earth observing satellite, airborne, and ground based remote sensing technologies. Scientific data is also being generated at an enormous rate due to climate, chemistry, biology, and nanotechnology. Earth observation (geospatial) and simulated (such as climate model derived) data are the harbinger of massive data inundation of the future. Clearly the success of scientific advancement and discovery will be strongly impacted by our capabilities in storing, analyzing, and creating meaningful information from the enormous databases within a potentially useful time frame. Although the progress of individual processor speed, cache performance, and graphics capability has been impressive, it has not been adequate to match the growth of available spatial data. This problem has been compounded by the drive towards real-time applications that require split-second response times to analysis on large and dynamic datasets (Xiong and Marble, 1996). The essence of this proposal is to address plausible high performance computational strategies for analyzing spatial geospatial data flow to provide accurate multi-sensor data analysis and registration for scientific, economic and policy analysis.

Today, spatial data have become an integral part of the decision making process in planning, policy, and operational missions for government agencies from local to national to global scales. Technological advancement resulting in both cost and time efficiency has prompted an explosion in the volume of spatial data that are being collected from remote sensors and developed from ground-based surveys (for example, high resolution imagery data from the Tennessee Base Mapping Program will exceed 3 terabytes (TB), and is only expected to multiply as the data are updated periodically). The data volume have been compounded by the generation of simulated data from high performance physical, chemical, and biological models and real-time data fluxes from in-situ sensor networks.

For time critical missions, high volume data and analysis must be provided on demand for addressing natural disasters (including floods, tornadoes, hurricanes, wildfires, diseases, and earthquakes) and deliberate attacks (including terrorist events, riots, and conventional warfare). To be effective the system must allow data inputs, access to a set of near-line simulation models, and visualization of observation and simulation data in real to near-real time. The simulation models will include those for hydrologic flow or runoff for floods and water-borne toxins, meso-scale weather simulations, wildfire, atmospheric dispersion model, transportation, epidemiological, interferometric SAR and GPS networks for earthquake, etc. The inputs of all of these diverse sets of data, models, and visualization environments will be geographically dispersed and must be linked through high performance networks to form an integrated presentation environment.



### 3. Participatory sensing and development of spatial data

Geographic data, describing objects and events, have been a fundamental component of scientific experiments and more importantly, in model calibration, verification, and validation for both physical and social sciences. The value of geospatial visualization of our environment and the increasing use of Geographic Information Systems (GIS) have also been well recognized and consequently spatial data have become critical to successfully addressing key issues such as good governance, poverty reduction strategies, and prosperity in social, economic, educational, and environmental health for government agencies from local to global scales. This realization has been mirrored with an increasing trend in the development of spatial data infrastructures by nations across local to state to national scales. Following the trend of spatial data infrastructure development, recent evolutions and advancements in geospatial and cyber technologies, combined with a population that is well informed and interested in global issues such as energy and climate, have cultivated an environment in which scientific research can potentially benefit significantly from the enormous volume of data that can be provided by citizens. Since 2009, the Open Government Data initiative has prompted a number of nations, including the United States and India, to commit to provide open access to government agency databases including geospatial information. Open access to non-sensitive government data may be perceived as an obligatory gesture to meet the general expectation of the general public but it is well realized that there are much broader benefits of empowering creative utilization of the open data for knowledge generation by academia, industry, government agencies, and individual citizens.

The ever-increasing demand of geospatial information has lead to an unprecedented rate of earth observation (geospatial) and simulated (such as climate model derived) data generation; often at the scale of terabytes to petabytes a day. Clearly the success of scientific advancement and discovery will be strongly impacted by our The critical challenge that faces the research and operational communities is, to the first order, develop and demonstrate capabilities in storing, analyzing, and creating meaningful information and applications from the enormous databases within a potentially useful timeframe. Secondly, to understand and assess the nature of geographic data produced by non-traditional sources such as individual citizens (also known as Volunteered Geographic Information or VGI) and its authenticity, validity, uncertainty and applicability in the context of spatial data infrastructure and sustainable development. Geographic Information Science (GIScience), including spatial database technologies, spatio-temporal data mining, high performance geocomputation, information retrieval, earth science informatics and knowledge discovery is a key area of research that are trying to address these critical scientific and technological challenges.

#### 4. Role of Geographic Information Systems

Traditionally, Geographic Information Systems (GIS) have been developed to address this data integration, analysis, and visualization issue. Past investments have only focused on developing spatial data infrastructures along with tools that are suitable for mega-scale data at best within desktop applications. In general, GIS functions are both involve intensive computing and I/O. However, recent research has focused much more on the former than the later (Healy et al., 1998). Complex data structures (raster and vector), with varying numbers of variable length data records spanning across linked files is typical in a GIS. Serial processing has been preferred over parallel to minimize large data pre-processing tasks. For these range of issues, ranging from database management to applied computational geometry, the task of developing a fully functional GIS in parallel environment has been daunting. However, advances in high performance computing provide ways to efficiently fill the research gaps and address utilization and user access of terascale data in a GIS.

#### 5. Data driven knowledge discovery

For knowledge discovery, characterization of the interaction between the human dynamics and transportation infrastructure is essential and requires integration of three distinct components, namely, data, models and computation. Recently, few models have started addressing the human dynamics of physical and social systems. However, none has been able to successfully integrate both the physical as well as behavioral aspects. Previous research, involving purely analytical techniques to simulations capturing microbehavior, has investigated questions and scenarios regarding the relationships among energy, emissions, air quality, and transportation. Primary limitations of past attempts have been availability of high resolution input data, useful “energy and behavior focused” models, validation data, and adequate computational capability that allows adequate understanding of the interdependencies of our transportation system. Progress has largely been limited by computational challenges necessary for accommodating the required high resolution along spatial, temporal and behavioral dimensions. This dimension is essential to characterize the interplay and interdependencies between (transportation) technologies and societal features that are likely to: (i) have an impact on the success of future technologies and (ii) be overlooked by current approaches of modeling at aggregated scales. To judiciously evaluate the impacts of multiple transformational mobility/energy/environment optimization strategies there is a clear need to create a modeling and simulation framework of regional transportation processes with high resolution geographic, demographic, socio-. economic data and behavioral characteristics. A limited number of parallel simulators exist today that scale efficiently to exploit high-performance computing platforms, can accommodate challenging combinations of large spatial regions (county and/or state level), fine time advances (minute by minute, for example) along long periods (weeks/months), and fine behaviors (e.g., individualized trip effects, vehicle emission effects,

group event effects, traffic controller induced congestion effects, etc.). Development of memory-full, non-linear fine-grained behaviors at the level of individual persons and fine process controllers, for a million or more individuals and/or physical devices, requires unprecedented support of power and efficiency from the underlying parallel simulation engines. To judiciously evaluate the impacts of multiple transformational mobility/energy/environment optimization strategies there is a clear need to create a modeling and simulation system of regional transportation processes with high resolution geographic, demographic, socio-economic data and behavioral characteristics.

## 6. Urban mobility and energy: A case study

At Oak Ridge National Laboratory (ORNL), we are combining the strengths of geospatial data sciences, high performance simulations, transportation planning, and vehicle and energy technology development to design and develop a national knowledge discovery framework to assist decision makers at all levels – local, state, regional, and federal. We have developed a modeling approach based on an individual consumer choice model that includes various socioeconomic variables defining sets of static and dynamic input to the model. Particular consideration was given to national data availability and scalability. This modeling and simulation capability allows national simulation of technology penetrations and their impact on climate (CO<sub>2</sub> emission) and electric energy infrastructures. In this spatially explicit model, we developed two novel concepts: a household synthesis model and a simulation of social diffusion of technology adoption using spatial proximity as one of the driving functions. The household synthesis model focused on investigating and developing a dependence-preserving approach in synthesizing household characteristics to support the activity-based traffic demand modeling. For the latter, the simple assumption was made that increasing exposure and awareness of new technology (alternative cars) with and without communication with spatial neighbors (for residents) and colleagues (at work) may provide a positive and a negative impact on potential adopters. Thus the simulation includes a flexible way to stipulate a distance threshold, which increases or decreases the likelihood of an individual adoption choice. The geographic scalability essentially describes the spatial extent of a particular phenomenon, in this case, the activities of a county's population, which in turn defines the volume and complexity of the data included in the simulation. Results from the simulation of Knox County, based on a 10% increase in first year PHEV adoption, shows that targeted adoption for families with annual income of \$60K and higher could impact 30% more vehicle miles traveled.

## Acknowledgement

This manuscript has been authored by employee(s) of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States

Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

## References

- [1] Bhaduri, B., E. Bright, P. Coleman, C. Liu, and J. Nutaro. 2009. "Integration of activity patterns through high resolution population distribution and dynamics." First Indo-U.S. Symposium on Mass Transit and Travel behavior (MTTBR 2009), Guwahati, India.
- [2] Bhaduri, B., C. Liu, J. Nutaro, and T. Zacharia. In press. "Ultrascale computing for emergency evacuation." In Wiley Handbook of Science and Technology for Homeland Security, J. Voller (Ed.).
- [3] Cui, X., Liu, C., Kim, H. K., Kao, S., Tuttle, M. A., and Bhaduri, B. L. 2010. A Multi Agent-Based Framework for Simulating Household PHEV Distribution and Electric Distribution Network Impact. To be presented at TRB 90th Annual Meeting, Washington D.C.
- [4] Dalal, P., C. Lui, and B. Bhaduri. 2008. "Developing trip chain sequence probabilities for a national household travel behavior model." Proceedings of the 71st American Pacific Coast Geographers Conference, Fairbanks, Alaska.
- [5] Healy, R., Dowers, S., Gittings, B., & Mineter, M., 1998. Parallel processing algorithms for GIS. Taylor & Francis, London. 460 pp. Xiong, D. & Marble, D. F., 1996. Strategies for real time spatial analysis using massively parallel SIMD computers: an application to urban traffic flow analysis. International Journal of GIS, 10 (6), 769-789.

## Applications of Micro Geo Data for Urban Monitoring

Yuki Akiyama (Research Fellow, Univ. of Tokyo, Japan)

### Abstract

Recently, various micro disaggregated data with spatially and temporally high resolution are being available which have high processability, for example detailed digital maps, mass person flow data and mobile census based on mobile phone GPS, and web information. We call such kind of big data the “Micro Geo Data (MGD)”. MGD have not been studied and have not used adequately even in academic research fields. On the other hand, specs of commercially available computers are improving, high-capacity hard disc drives are becoming widely used and high-performance GIS software and many kinds of open GIS software are being available these days. In addition, many kinds of MGD are being opened to the public by some Japanese local governments or being commercialized by private companies. Because of this situation, studies and utilizations of MGD enable to develop new research fields and to realize new researches which were physically impossible previously. It is widely expected to develop new research fields which could not realize by previous studies by proper utilization of MGD. Therefore, this paper introduces various MGD to be on the verge of becoming possible to be available in Japan and our studies to use MGD and to develop new MGD for urban monitoring. This paper introduces studies to monitor various phenomena in macro scale areas i.e. broad urban areas or national land of Japan using many kinds of MGD, e.g. time-series detailed maps and telephone directory, web information, and mass person flow data, and so on. In addition, this paper also introduces our ongoing studies to develop new MGD and visualize MGD. It is expected that MGD become widespread and become widely used in various fields, e.g. academic researches, operations by local government or business activities by private company. Researchers should catch up methods of handling and utilization of such new spatial and temporal data and statistics and prepare to be able to accept requests from the world. For this request, we should acquire and share broad-based knowledge of researches and operations to utilize MGD today or that have future availability of MGD.

**Keywords—** Micro geo data (MGD), Big data, Urban monitoring, Time-series data, Visualization

## 1. INTRODUCTION

Recently, performance of computers and their peripheral devices is improving, their prices are reducing and internet environment are developing rapidly. We are accessible for high resolution digital maps and satellite images which used to be utilized by some researchers and strategists because of them. In addition, person flow big data have received a lot of attention in recent years e.g. person flow data collected by mobile phone GPS and mobile statistics developed by them. Furthermore an enormous amount of information is being accumulated on the web every day.

Because of this situation, an enormous amount of information, i.e. the “Big data” is becoming accessible for us today. In Japan, many attempts to use various big data in meaningful ways have started in various fields of industry, government and academia. In the case of fields of urban and regional analysis, it is expected to flourish researches to utilize geospatial and geotemporal big data, i.e. the “Micro Geo Data (MGD)” in the near future. It is on the verge of becoming possible to acquire large quantity of disaggregated spatio-temporal data and realize detailed urban and regional monitoring by adequate acquisition and processing of MGD. This paper introduces study cases related to urban monitoring to use MGD in Japan. In addition, it introduces new visualization tool of MGD. Finally, we introduce challenges of MGD utilization.

## 2. URBAN MONITORING USING MGD

### 2.1 Micro scale time-series changes in Urban area – Visualization of time-series changes of all shops and offices

Micro scale time-series changes, i.e. time-series changes of each shop and office are interesting information for understanding of urban transformation. In Japan, it can be monitored using digital telephone directory. Digital telephone directory is digitized data of telephone directory throughout Japan. Japanese telephone directory contains detailed information of each shop and office, e.g. shop and office name, their business categories and detailed locations: their addresses and occupied floors and room numbers. In addition, we can monitor time-series changes of each shop and office, i.e. continuation, change, emergence and demolition of them between two different years using them in two different years (Akiyama et al. [1]).

Fig. 1 shows a visualized result of time-series changes of each shop and office to integrate all shops and offices spatially of telephone directory in 2003 and 2008 and to identify their names automatically. In addition, activity of time-series changes of all shops and offices can be quantified and visualized to aggregate this data as shown in Fig. 2. Changed rate of shops and offices are defined by the method as shown in Fig. 3 and briskness of



commercial activity is visualized. This data can be updated frequently because Japanese telephone directories are updated every 2 months. Time-series changes of every shops and offices in arbitrary area and in arbitrary two different years can be monitored with high-frequency updating.

## 2.2 Mass person flow data

Many previous studies monitored dynamic population and number of visitors in specific commercial areas or facilities in urban areas by field survey and questionnaire survey commonly. Information accumulated these method have relatively high reliability. However, it is difficult to conduct same kind of surveys frequently in broad area because of large amount of labor and time for these methods of survey.

For this problem, extensive questionnaire surveys called the “Person trip survey” have been conducted in Japan by Japanese government. This survey collected personal attributes, e.g. age, gender, their home and work locations and their visiting locations and transportation devices of randomly-selected persons in one day. In the case of latest person trip survey in the Greater Tokyo

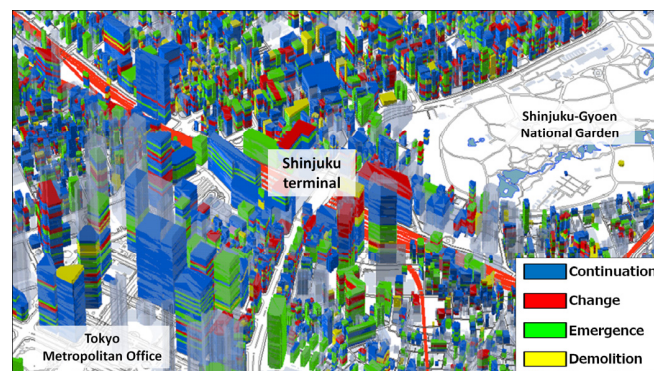


Fig.1 3D Time-series changes of shops and offices between 2003 and 2008 in the Tokyo metropolitan area

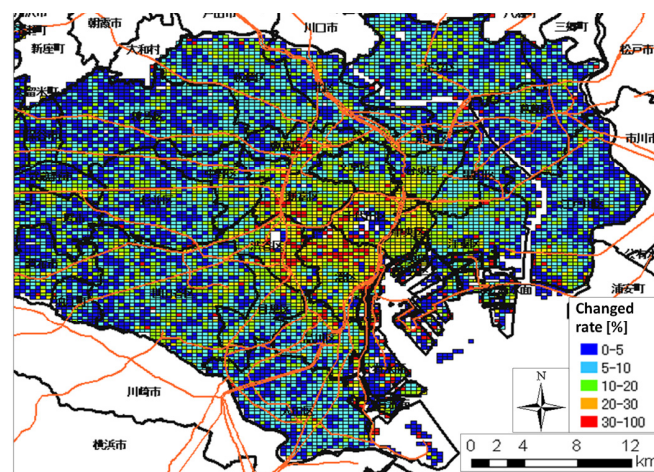


Fig. 2 Changed rate of shops and offices in the center of Tokyo between 2003 and 2008 accumulated by 500m square grid

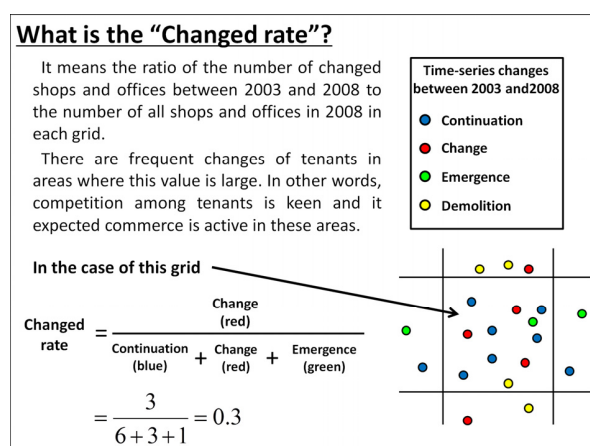


Fig. 3 Definition of the changed rate

Region in 2008, person trip information of about 340 thousands households was collected. In addition, person distributions can be estimated in arbitrary time i.e. every 1 minute in one day by spatial interpolation of route using detailed road and railway network data. This data is called the “People flow data” [2]. Fig. 4 shows people distribution with transportation modes in AM8:00, 2008 by the people flow data.

Furthermore, utilization of mobile phone GPS data has received attention in recent years. Some studies have already tried to monitor person flows in broad area using location data collected from GPS devices. In addition, mass person flows are being able to monitored everyday throughout Japan using mass GPS data collected by mobile phone applications with permission to use them from mobile phone users (Sekimoto et al. [3]).

Mass GPS data can monitor areas where persons stayed in arbitrary times and areas. Fig. 5 shows estimated number of visitors in each commercial area in the center of Tokyo to integrate staying areas extracted from annual mobile phone GPS data in 2012 with the Commercial accumulation statistics which is introduced below (Akiyama et al. [4]). Fig. 6 shows estimated average time-series number of visitors at some main commercial areas in Tokyo in weekday and the weekend. Same kind of result in specific day can be developed. In the future, we plan to analysis relation between the number of visitors in commercial areas with effects of weather, accidents and events, and so on.

## 2.3 Gathering of shop and office information from the Web

Web API service is the API to use web services. Main search engines e.g. Google, Yahoo, map search services e.g. Google maps, Bing maps and various websites of companies introduce this service. Using this system, we can gather search results by search engines, maps in arbitrary areas, merchandise information and shop information, and so on. Fig. 7 shows a map to integrate digital map with shop information gathered from the Hotpepepr API. The Hotpepper is one of the most popular portal sites about gourmet in Japan. Though collectable number of shops by the API is less than telephone directory or digital maps, this data is

fresher and contains richer information e.g. the number of seats, the availability of parking, business hours, and so on than them.

Furthermore, we can gather fruitful information about shops to analyze information of web sites searched by search engines. Fig. 8 shows an example to gather business hours of each shop to gather related websites by an API service of search engine and analyze their html (Okamoto et al. [5]). Search phrases are name and address of each shop collected from digital telephone directory. In addition, Fig. 9 shows time-series change of open shop distributions in the center of Tokyo. We can monitor daily dynamic changes of Tokyo to collect business hours of about 500 thousands shops.

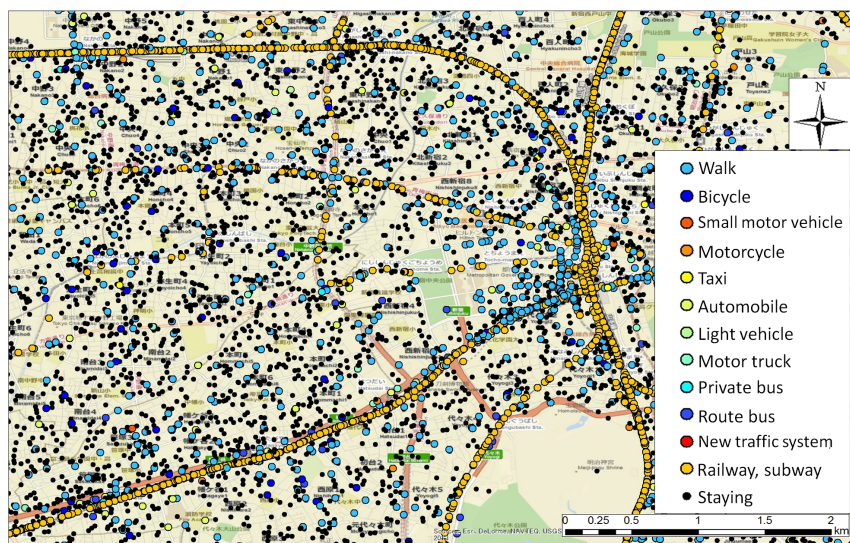


Fig. 4 People distribution with traffic mode in the center of Tokyo by the people flow data (AM8:00, 2008)

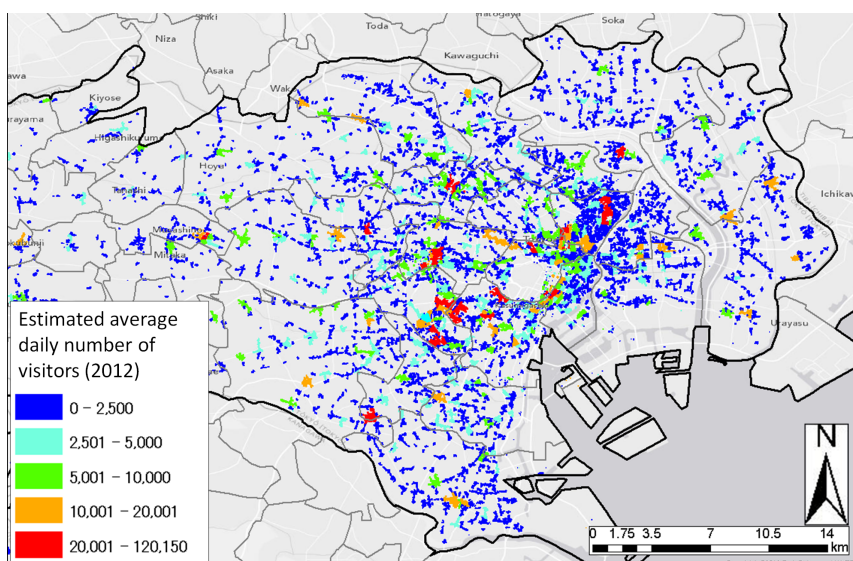


Fig. 5 Estimation of average daily number of visitors in each commercial area in the center of Tokyo in 2012



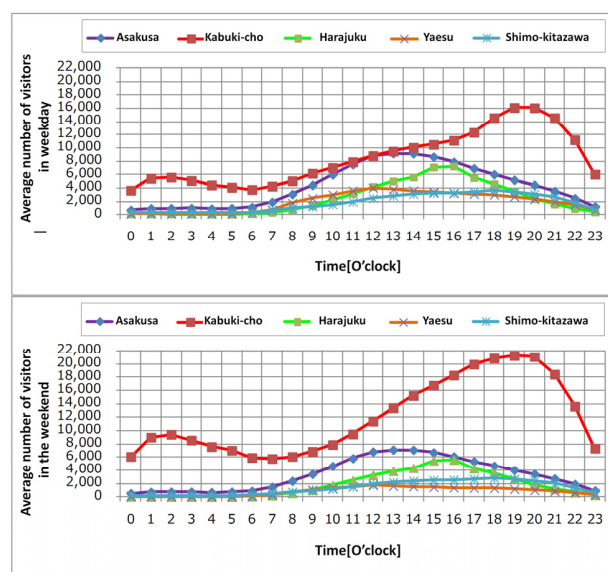


Fig. 6 Time-series changes of the average number of visitors in some main commercial areas in Tokyo in weekday and the weekend

## 2.4 Detection of the “hot area” using search results by search engines

The number of hit by specific search phrase can be gathered using API services of search engines. It is expected that areas or shops which are searched by web search engines frequently are popular in reality space where have attracted attention in recent years and gather many people. We call such area the “Hot area”. Fig. 10 shows distributions of hot areas in Setagaya Ward, Tokyo based on this hypothesis (Akiyama et al. [6]). First, shop names and addresses in Setagaya Ward were collected from digital telephone directory. Second, numbers of hit of each shop were collected by API of a search engine based on each name and address. Finally, numbers of hit of each shop were accumulated into 250m square grids. Areas where height of grid is large were searched many times. In addition, areas where the number of hits per shop was large are colored orange to red. If these both values are large, the area is the hot area.

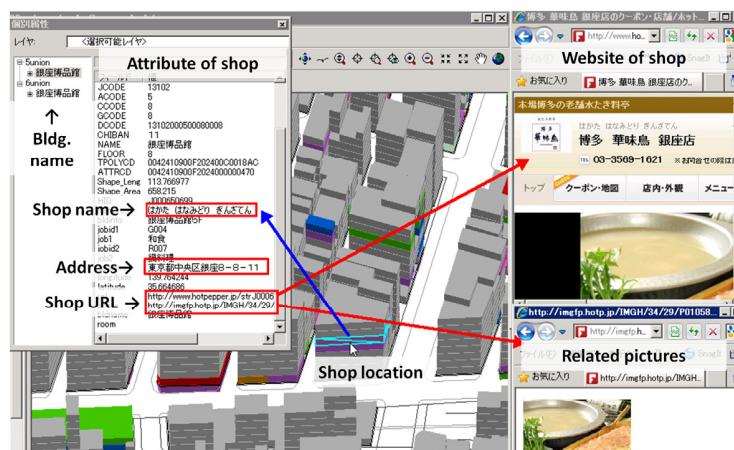


Fig. 7 Visualization of shop information gathered by a Web API service (in the case of Ginza, Chuo-ku, Tokyo in 2011)

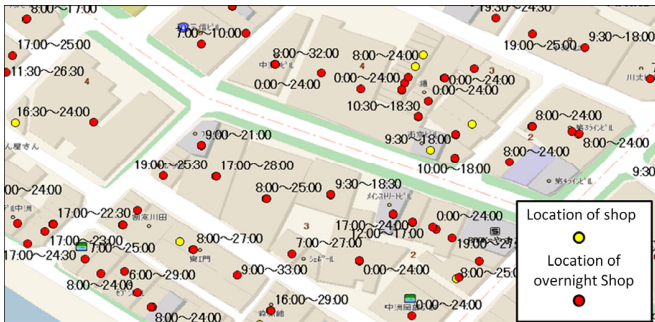


Fig. 8 Business hours of shop collected by analytical results  
 ated web page html(in the case of Nakasu, Fukuoka city in 2012)

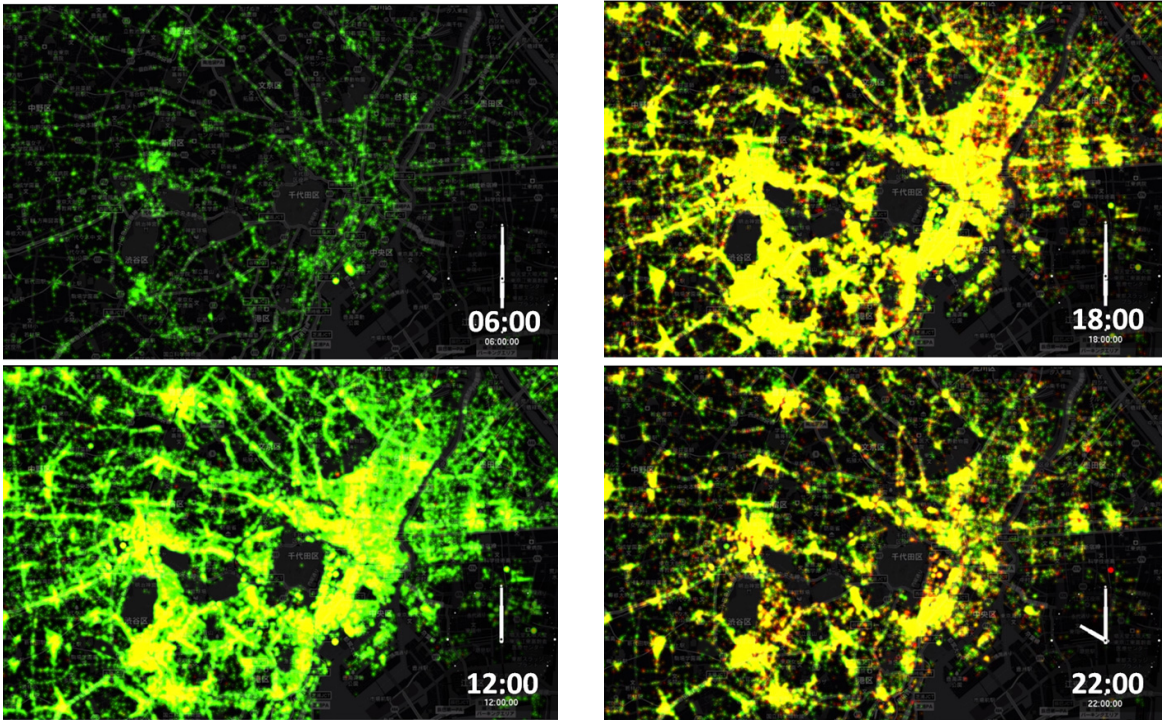


Fig. 9 Time-series distribution map of open shops in the city center of Tokyo in 2012

### 3. DEVELOPMENT OF NEW MGD

### 3.1 Commercial accumulation statistics

It is very interesting research subjects to develop new MGD to be helpful in urban monitoring. The “Commercial accumulation statistics” is such a new MGD.

The commercial accumulation statistics is the polygon data to be able to monitor locations of commercial area throughout Japan (Akiyama et al. [7]). It is the first data in Japan to be able to monitor not only locations of commercial areas but also their spatial extent. In addition, the data can monitor the number of shops on each business category, the rate of chain shops, areas, and so on of each commercial area.

The commercial accumulation statistics was developed using location data of shops and offices collected from digital telephone directories. First, shop and office data to construct commercial areas are selected from telephone directories and point data of shop and office are developed. Second, buffer polygons which construct commercial areas are created by our original spatial processing from each shop and office points. Finally, polygons which express spatial extents of commercial area were developed to integrate with overlapping buffer polygons. This method realizes to develop buffer polygons of commercial accumulations which their locations and spatial extent are similar to our actual feeling to calculate average distances between each shop in each commercial area automatically. We have already realized to develop this data throughout Japan as shown in Fig 12 to apply this method to Japanese telephone directory which contains about 10 million shops and offices.

We can monitor actual states and time-series changes of commercial accumulations from various viewpoints to use this data. For example, Fig 11 shows a map to overlay polygon data of commercial accumulations in 2007 and 2011 in Aomori city, Japan. Aomori city is a local city with approximately 300 thousand populations in northeastern part of Japan. We can find some depressed and emerging commercial areas in outskirts. On the other hand, there are some depressed commercial areas in the city center of Aomori city i.e. districts in front of Aomori terminal.

This data has already been utilized in some studies and it is provided from the JoRAS: Joint Research Application System by CSIS: Center for Spatial Information Science, the University of Tokyo for research aims. In addition, this data was commercialized as marketing contents from a joint research company [8].

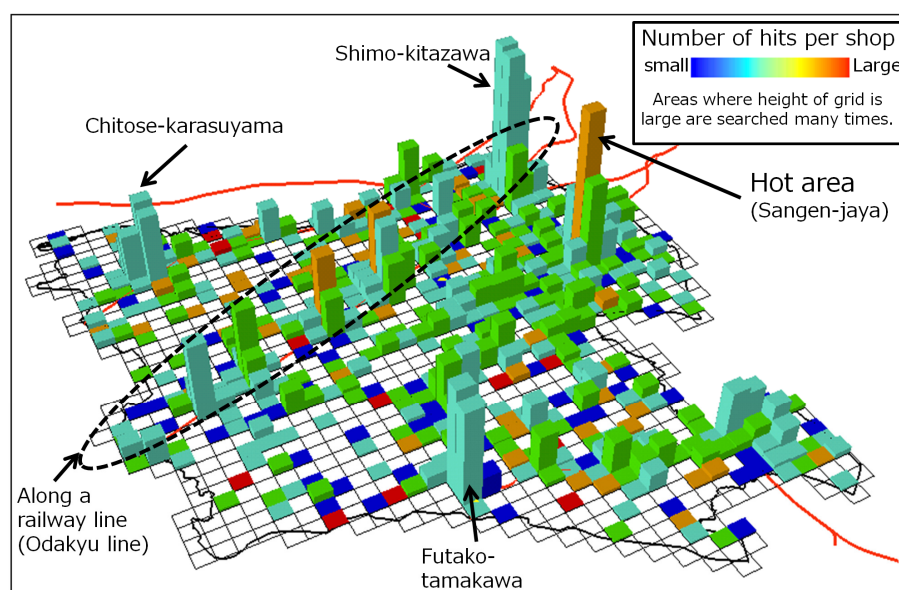


Fig. 10 Detection of hot areas based on numbers of hit of each shop collected by an API service of search engine (in the case of Setagaya Ward, Tokyo in 2009)



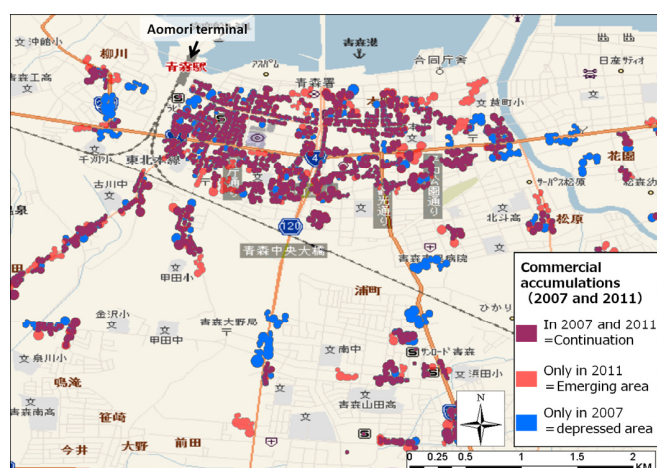


Fig. 11 Time-series distributional changes of commercial accumulation in Aomori city in 2007 and 2011

### 3.2 Micro population census

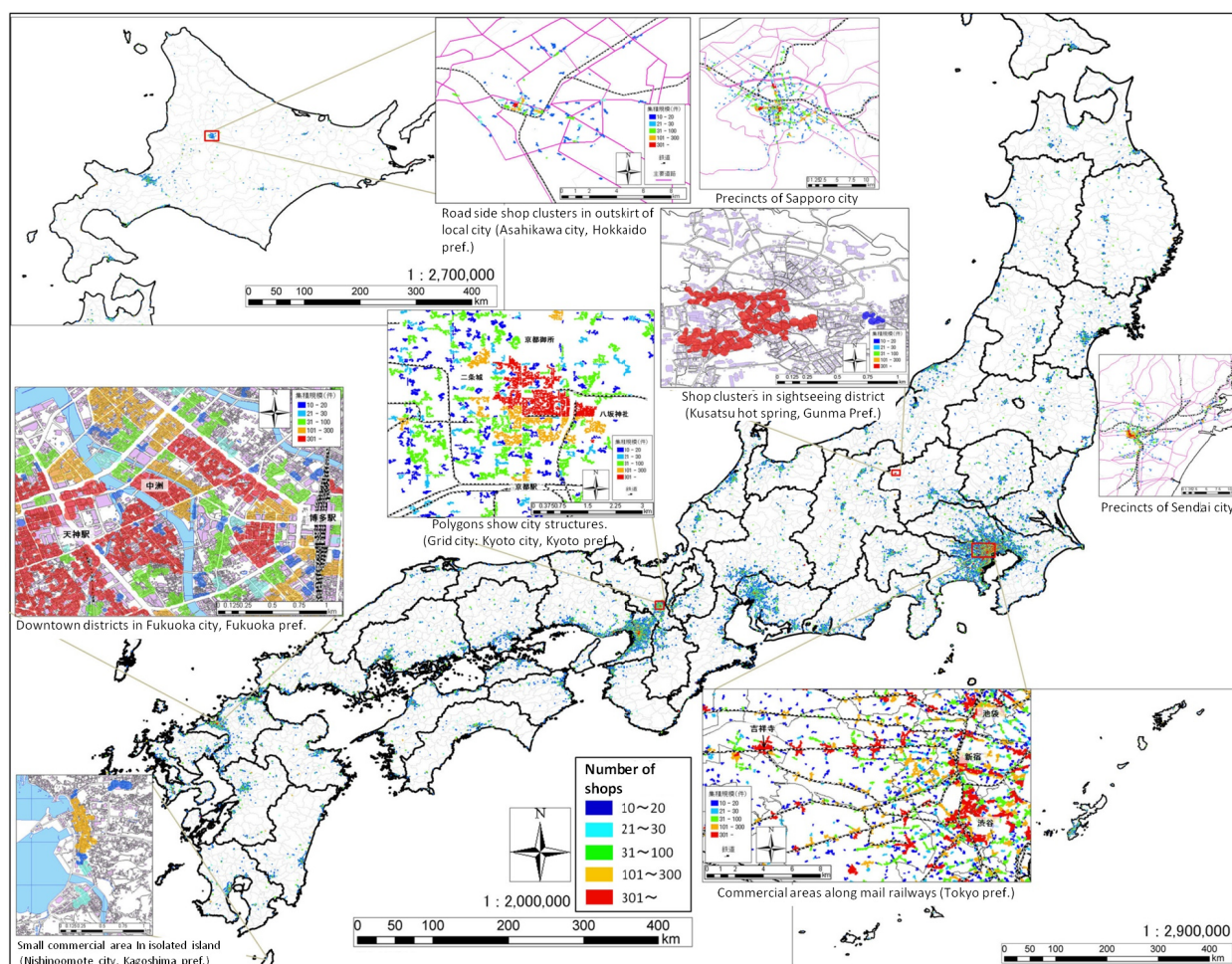


Fig. 12 Commercial accumulation statistics throughout Japan in 2009

In Japan, the population census has been used for monitoring distribution and movement of population. However, it is difficult to monitor detailed distribution of population because spatial units of open population census are 1km or 500m square grids or administrative unit, i.e. city, ward, town and village. In addition, there is a problem that population is evenly-distributed even in areas where population is eccentrically-located to aggregate into grid or administrative unit. Recently, there is a needed for detailed information about population distribution with high reliability in the fields of regional analysis for disaster management, traffic planning, pinpoint marketing, and so on. In spite of this situation, these limitations are highly influential for theme aims.

Therefore, we developed the “Micro population census” which is estimated distribution data of household and resident of Japan (Akiyama et al. [9]). This data realized to distribute information about household and resident from some charts of the population census in locations where households are located collected from detailed digital maps of Japan by combination with other various statistics. This data is sort of the “artificially disaggregated population census” and new population census which can be aggregated into arbitrary spatial unit. Even though disaggregated points of this data do not necessarily coincide with actual states, actual states of population can be reproduced to integrate the data into spatial units e.g. city blocks or grids. Fig. 13 shows disaggregated micro population census and aggregated result by city blocks. In addition, Fig. 14 shows the aging rate aggregated into 250m square grid. This data have been also used some researches and we also use this data to estimate human suffering by heavy earthquakes throughout Japan (Akiyama et al. [10]; Ogawa et al. [11]).

### 3. VISUALIZATION OF MGD

MGD is new big data with high resolution spatially and temporary. Therefore, it is difficult to visualize and analyze MGD using existing GIS software. Visualization and analysis of person flow data is especially major problem because such kind of data are constructed by huge number of staying and flowing points of each person. The person flow data in this paper are enormous number of GPS data from mobile phones which are becoming available recently, SNS data with geo tags and the person trip survey, and so on. The person trip survey is massive questionnaire survey for randomized households to monitor people flow and stay in major metropolitan area of Japan.

Person flow data are constructed by stay and flow locations and times of each person. They are visualized by common existing GIS software “statically”. However, there are few GIS software to be able to visualize “dynamically” and analysis them spatio-temporally. Previously, we needed to develop dedicated applications for efficient processing of time-series spatial data e.g. visualization of person movement locus, detection of events.

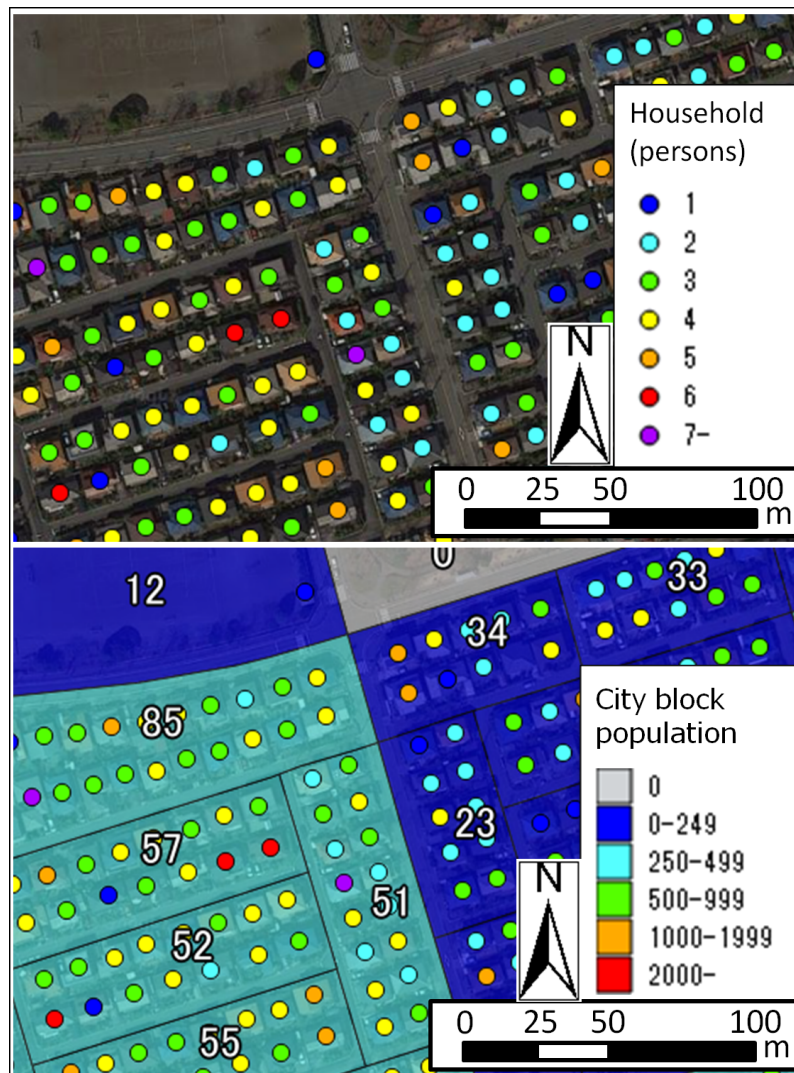


Fig. 13 Disaggregated micro population census and aggregated result

Therefore, new GIS software called the “Mobmap” is being developed for the purpose of visualization and analysis of person flow data (Ueyama, [12]). Fig. 15 shows a screen shot of animations of time-series person flows on digital map using the Mobmap. A person flow data in this figure is the “SNS-based People Flow Data” (Nightley, Inc. [13]). It is realized that locations of persons collected from SNS data with geotag are interpolated simply by their moving paths. By using this function, the Mobmap can capture and count persons who pass certain road and move in and out certain area. In addition, various functions of spatial analysis are being developing and implemented to the Mobmap. There are being some efforts to make it easier for someone to visualize and analysis of MGD than before in this way.



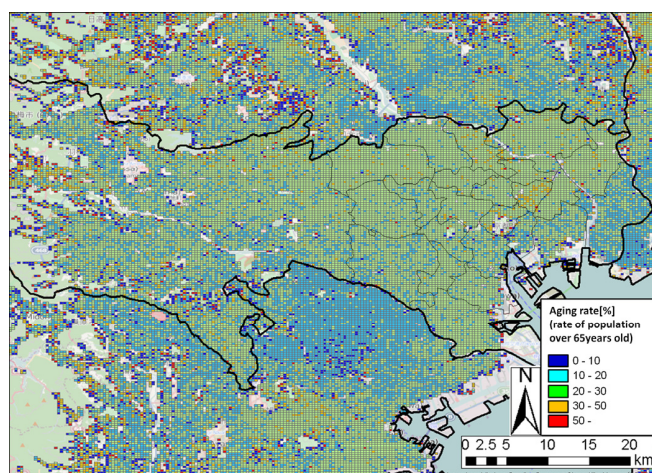


Fig. 14 Aging rate in Tokyo aggregated into 250m square grid in 2005

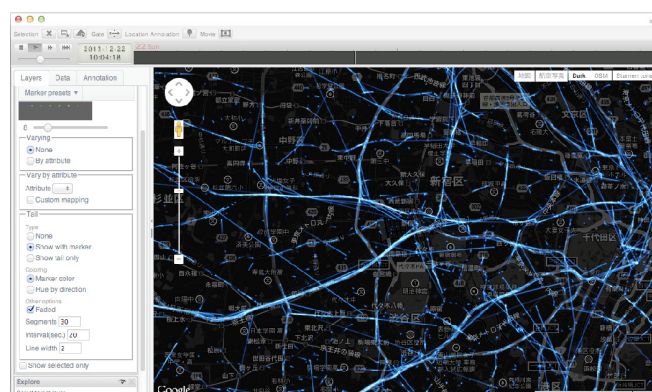


Fig. 15 Screen shot of animations of SNS-based time-series person flows in Tokyo by the Mobmap

## 5. CONCLUSION

Recently, various methods for urban monitoring using MGD are coming true. In addition, high-resolution big data with locations and times throughout Japan or world, namely “MGD” are being accumulated and updated every day which is almost impossible to accumulate and handle before. We think utilization of MGD will become one of the important technologies in the field of urban monitoring and analysis in the near future. Researchers in these fields should acquire techniques and knowledge to handle and interpret MGD for such a future.

On the other hand, it is believed that there is information which cannot be explained adequately only from data or information which can be collected by field surveys even though various MGD are available. Therefore, it is considered that researchers will be required to acquire techniques and sense to use MGD and field data properly or to integrate them.

In Japan, now is the stage to finally start full-scale utilization and application of MGD as this paper introduces. We think that we should discuss more fruitful methods for utilization of MGD to improve the world than now with researchers and experts in various fields. We therefore launched the “Micro Geo Data Forum” for this aim and are

accumulating, developing and promoting MGD [14]. If you are interested in this activity or MGD introduced in this paper, please contact the author.

## Speaker Infomation

Yuki Akiyama is an Assistant professor of Earth Observation Data Integration and Fusion Research Institute, The University of Tokyo and a visiting researcher of Center for Spatial Information Science, The University of Tokyo. Contact: Institute of Industrial Science, Cw-503 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan (phone: +81-3-5452-6417; fax: +81-3-5452-6414; e-mail: aki@iis.u-tokyo.ac.jp)

## Acknowledgment

Author was given some MGD by the joint research with Center for Spatial Information Science (CSIS), The University of Tokyo (research ID: 122) and ZENRIN Co., LTD. Author was also supported by Earth Observation Data Integration and Fusion Research Institute (EDITORIA). In addition, author was supported by Korea Research Institute for Human Settlements (KRIHS) that gave an opportunity to make this presentation and bear all travel expenses. I would like to thank CSIS, Zenrin, EDITORIA and KRIHS for their considerable contributions.

## References

- [1] Y, Akiyama. and R, Shibasaki., “Spatio-temporal Integration Method for Shop and Office Data with Location Information and Application for Urban and Regional Analysis”, Theory and applications of GIS, 19(2), 2011, pp.57-67. (in Japanese)
- [2] People flow project, <http://pflow.csis.u-tokyo.ac.jp/index.html>, Last access date: August 13, 2014.
- [3] Y, Sekimoto., H, Teerayut. and R, Shibasaki., “Trend of People Flow Analysis Technology Using Mobile Phone”, Information processing, 52(12), 2011, pp.1522-1530. (in Japanese)
- [4] Y, Akiyama., H, Teerayut. and R, Shibasaki., “Time-series Analysis of Visitors in Commercial Areas Using Mass Person Trip Data”, Proceeding of 22nd GISA annual conference, 2013, C-5-4. (in Japanese)
- [5] Y, Okamoto., Y, Akiyama., S, Ueyama. and R, Shibasaki., “Visualization of Business Hours for Shops and Offices Classified by Business Categories in Shopping Area”, Proceeding of The 32nd Asian Conference on Remote Sensing, 2011, TS3-10.
- [6] Y, Akiyama., H, Sengoku. and R, Shibasaki., “An Attempt of Regional Analysis Using Web Search Result of Geographic Information”, 2009 Korea Japan GIS International Symposium Joint Conference of KAGIS Fall Symposium, 2009, pp.84-87.

- [7] Y, Akiyama., H, Sengoku., H, Takada. and R, Shibasaki, “Development of Commercial Accumulation Polygon Data Throughout Japan Based on the Digital Classified Telephone Directory”, Proceeding of CUPUM2011, 2011, F-TC-3(1).
- [8] ZENRIN Co., LTD., “Marketing contents – Commercial accumulation statistics 2013–”, <http://www.zenrin.co.jp/product/gis/marketing/marketing02.html>, Last access date: August 13, 2014. (in Japanese)
- [9] Y, Akiyama., H, Takada. and R, Shibasaki, “Development of Micropopulation Census through Disaggregation of National Population Census”, CUPUM2013 conference papers, 2013, 110.
- [10] Y, Akiyama., Y, Ogawa., H, Sengoku., R, Shibasaki. And T, Kato., “Development of Micro Geo Data for Evaluation of Disaster Risk and Readiness by Large-scale Earthquakes Throughout Japan”, Proceeding of Annual conference on Infrastructure and Management, 2013, 392. (in Japanese)
- [11] Y, Ogawa., Y, Akiyama. and R, Shibasaki, “The Development of Method to Evaluate the damage of Earthquake Disaster Considering Community-based Emergency Response Throughout Japan”, GI4DM2013, 2013, TS03-1.
- [12] S, Ueyama, “mobmap for Chrome”, <http://shiba.iis.u-tokyo.ac.jp/member/ueyama/mm/>, last access date: August 13, 2014.
- [13] Nightley, Inc., “SNS-based People Flow Data”, <http://nightley.jp/archives/1954>, last access date: August 13, 2014.
- [14] Micro Geo Data Forum, <http://geodata.csis.u-tokyo.ac.jp/>, last access date: August 13, 2014.



## 3

## How to use Big Data in LH

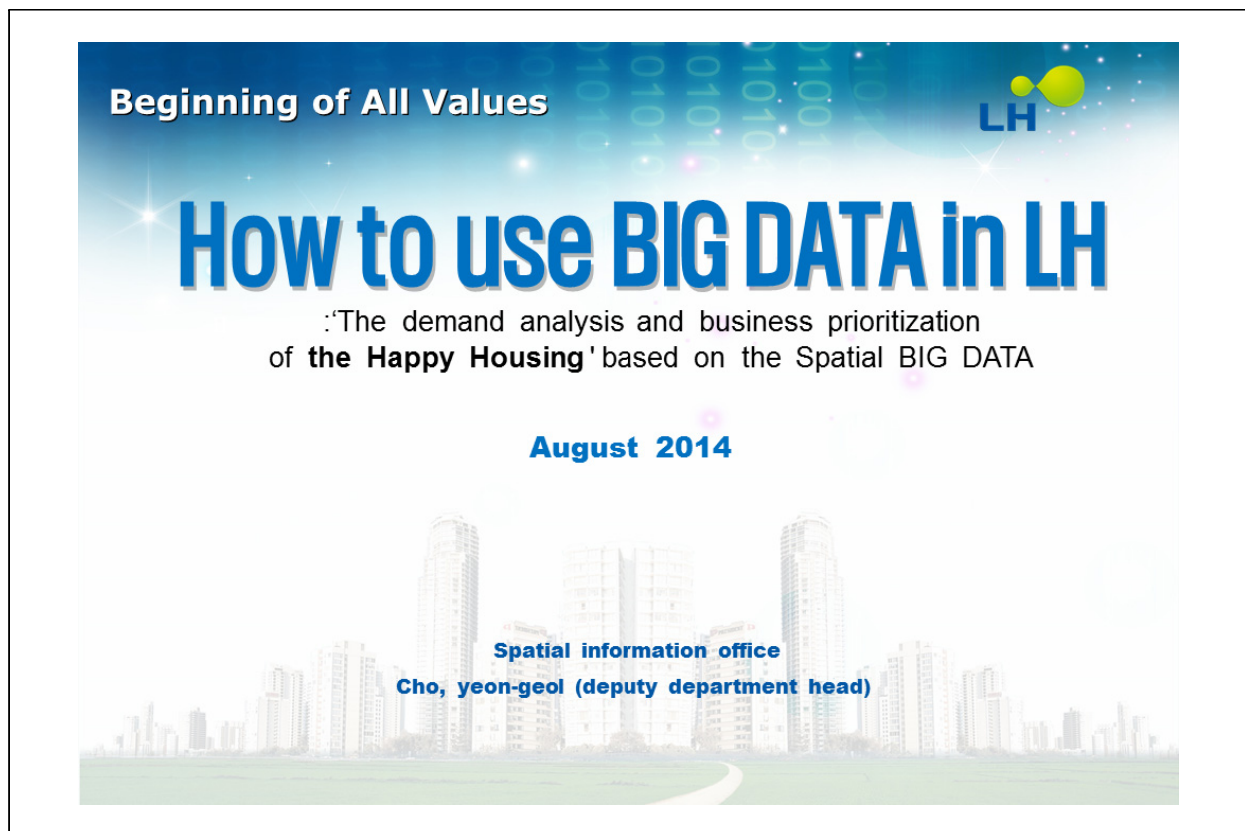
Yeon-Gurl Cho

(Vice Director of Spatial Information Division, Korea Land &amp; Housing Corporation)

## ABSTRACT

'The demand analysis and business prioritization of the Happy Housing' based on the Spatial BIG DATA

- Introducing the 'Happy Housing', a form of constructed rental housing by LH for the class of newly married couple, university student and someone who is just starting out in a career etc.
- The demand analysis of the Happy Housing by class and that of the spatial visualization



## PRESENTATION FLOW

1. About LH
2. About the 'Happy Housing'
3. How to use Big Data in LH
  - (1) Why do LH use Big Data for analysis?
  - (2) LH Big Data summary
  - (3) Comparison analysis
  - (4) Improvements
  - (5) The class segmentation of the 'Happy housing'
  - (6) The distribution of each class
  - (7) Modification
  - (8) Conclusion(Example)

### • About LH

## ▶ LH is . . . . .

### LAND & HOUSING CORPORATION

\* LH, a state-owned enterprise, was founded with a vision of providing affordable and quality public housing—the key to national happiness—and leading efficient land development in order to enrich the lives and living conditions of Korea people.

- To construct and supply decent and affordable housing units
- To develop housing land, new towns, Multi-functional Administrative City and so on
- To develop industrial and logistics complexes(ex. Kaesong Industrial complex) and overseas land
- To perform land reserve and management, rental housing management, & land and housing informatization

## • About the 'Happy Housing'

### the 'Happy Housing'

► - the 'Happy Housing', a form of constructed rental housing by LH for the class of newly married couple, university student and someone who is just starting out in a career etc.



## • how to use Big Data in LH

### ► Why do LH use Big Data for analysis?

#### Why... Big Data.....?

Big Data can show the way to solve the problems more efficiently based on large and various data.

This analysis is made for more efficient decision making in 'Happy Housing project' than before.



## • how to use Big Data in LH

### LH Big Data...

- ▶ **What**: The demand analysis of the 'Happy Housing' by class and that of the spatial visualization
- ▶ **Why**: More efficient demand analysis for business prioritization of the Happy Housing'
- ▶ **How**: Convergence of governmental and social data

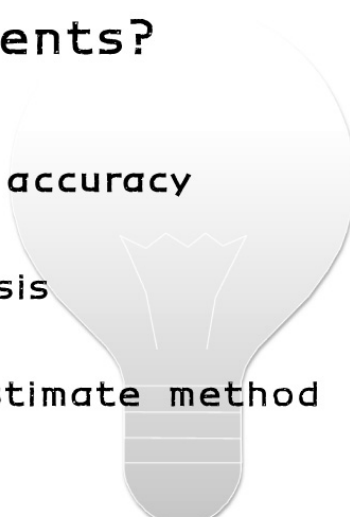
## • how to use Big Data in LH

### ▶ Comparison analysis

	before	after
<b>Volume of data</b>	Small data No convergence of data	Big data Convergence of data
<b>Latest data</b>	Mostly past data	Latest data
<b>Origin of data</b>	Inside data (Governmental data)	Including outside data (Social data)
<b>Method of investigation</b>	Sampling method	Total inspection method
<b>Standard of probe region</b>	Address	Including real residence (Modification : from address to real residence by communication data)

• how to use Big Data in LH

Improvements

- ▶ What are improvements?
  - ▶ Improvement of analysis accuracy
  - ▶ Various and speedy analysis
  - ▶ More accurate demand estimate method
- 

• how to use Big Data in LH

The class segmentation

- ▶ The class segmentation to meet the requirements of the 'Happy Housing'

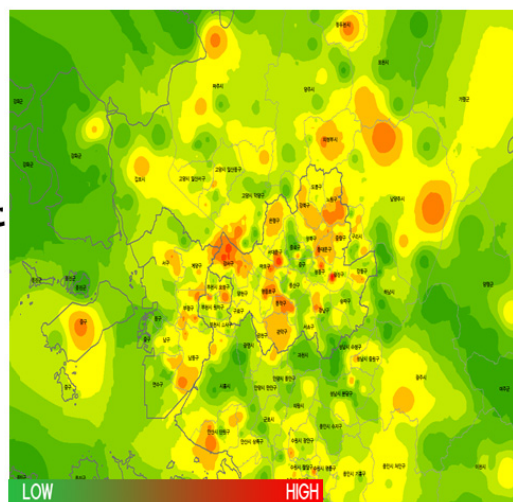
The class segmentation				
university student	newly married couple	someone who is just starting out in a career	Low-income group	Senior citizen households



- how to use Big Data in LH

Distribution

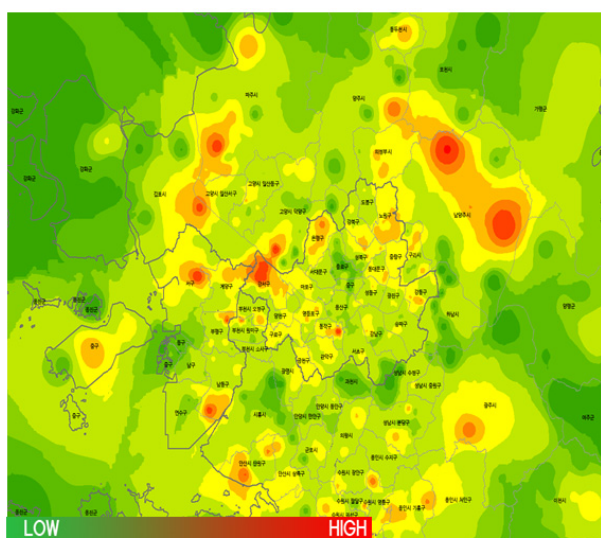
► The distribution of university student around the Capital area



- how to use Big Data in LH

Distribution

► The distribution of newly married couple around the Capital area

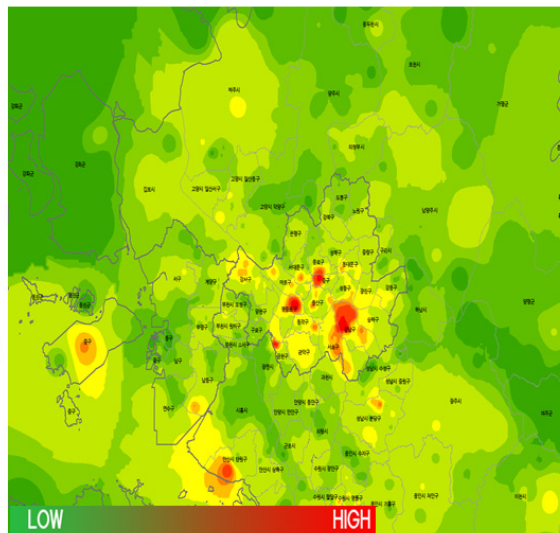




- how to use Big Data in LH

### Distribution

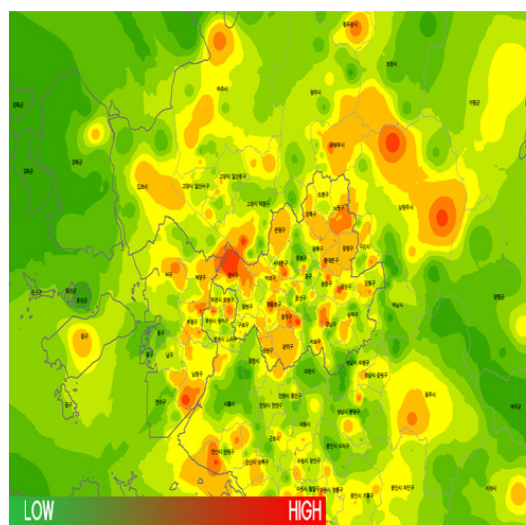
► The distribution of someone who is just starting out in a career around the Capital area



- how to use Big Data in LH

### Distribution

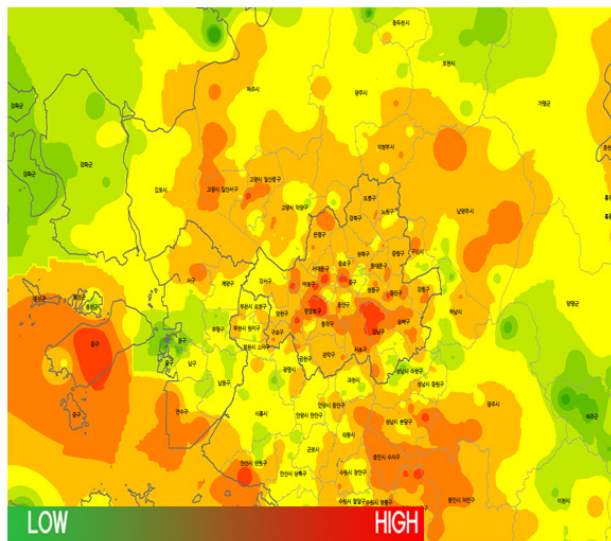
► The distribution of Low-income group around the Capital area



- how to use Big Data in LH

Distribution

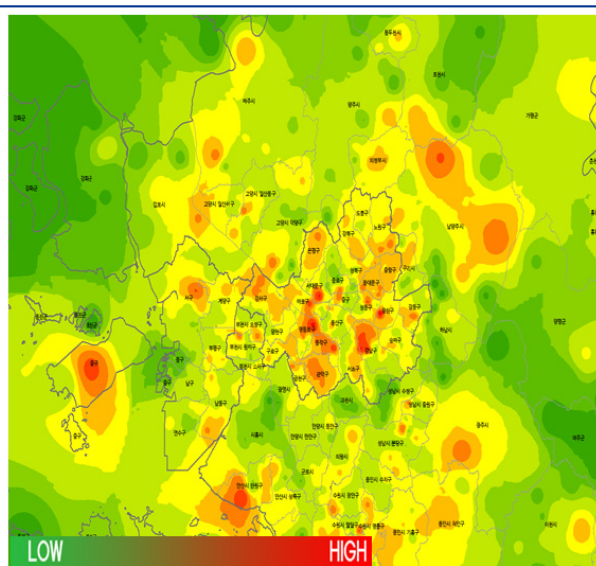
► The distribution of Senior citizen households around the Capital area



- how to use Big Data in LH

Distribution

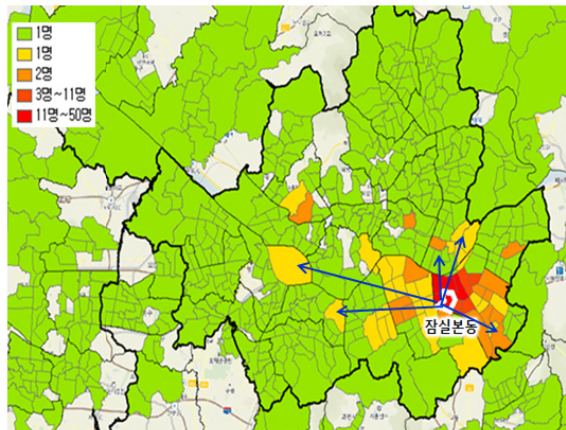
► The distribution of all classes around the Capital area



## • how to use Big Data in LH

### Modification

▶ **Modification**  
(From address  
to real residence  
by communication  
data)

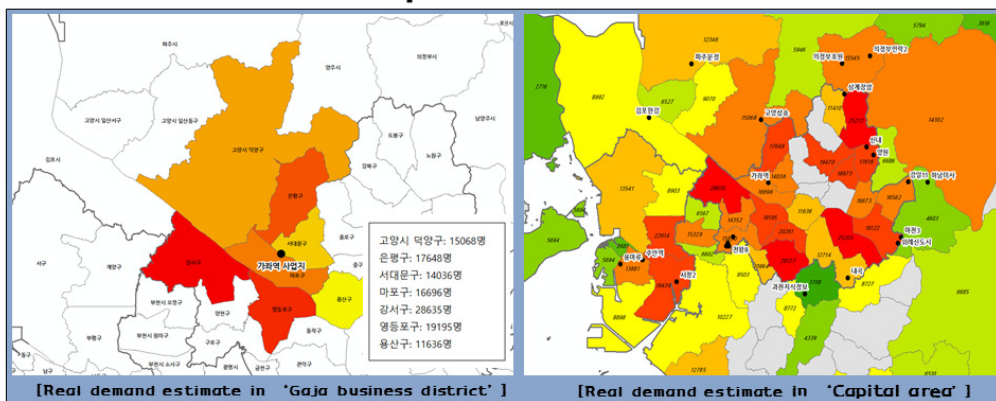


[Real residence of people that their address is 'JAMSIL BONDONG']

## • how to use Big Data in LH

### Conclusion (Example)

▶ **Real demand estimate in 'Gaja business district' and 'capital area'**



[Real demand estimate in 'Gaja business district']

[Real demand estimate in 'Capital area']

