Data Markeplace Initiative

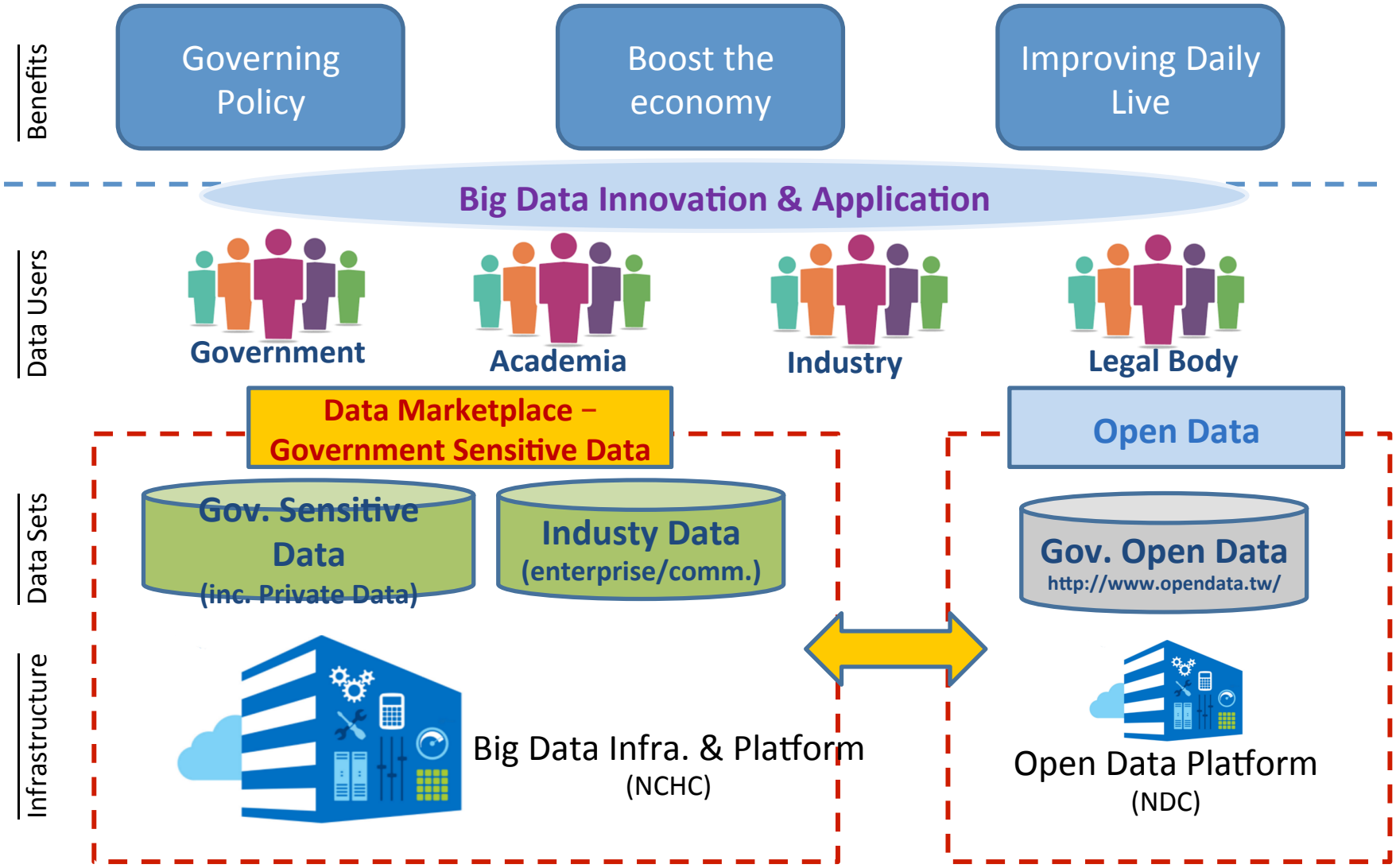
Weicheng Huang
National Center for High-performance Computing
National Applied Research Laboratories,
(NCHC), Taiwan

Contents

- Goal
- Activity
  - Legal obstacles
  - mechanism
  - Tool/Software
  - Platform & Service

Vision of Data Mart

- Data Economy requires the application of “sensitive” data.
- Secured Data Platform shared by the government and industry, can boost the data value



Data Marketplace

1 Data Cloud & Data Integration

- **Data backup**：hosting important/sensitive data w/o utilizating cold data
- **Large scale computing**：Big Data Analysis Platform, e.g. Hadoop, … etc.

2 Tech. & efficiency

- **Variety of Data Analysis tools**：big data toolkits, data processing tool, … etc. Tool/tech. for Big Data **Applications**：promoting Big Data tools to gov. agencies

3 Data Analysis & Vis.

- **Data Processing, promote data value**：data cleaning, catalogue, aggregation, … etc. work with the Data Analysis.
- **On-line visual analysis tool & env.**：selection and adoption of tools/software

Benefiting Gov. Agencies

Security & low cost

- Reduce the cost of hardware/software
- Shared data center & data management

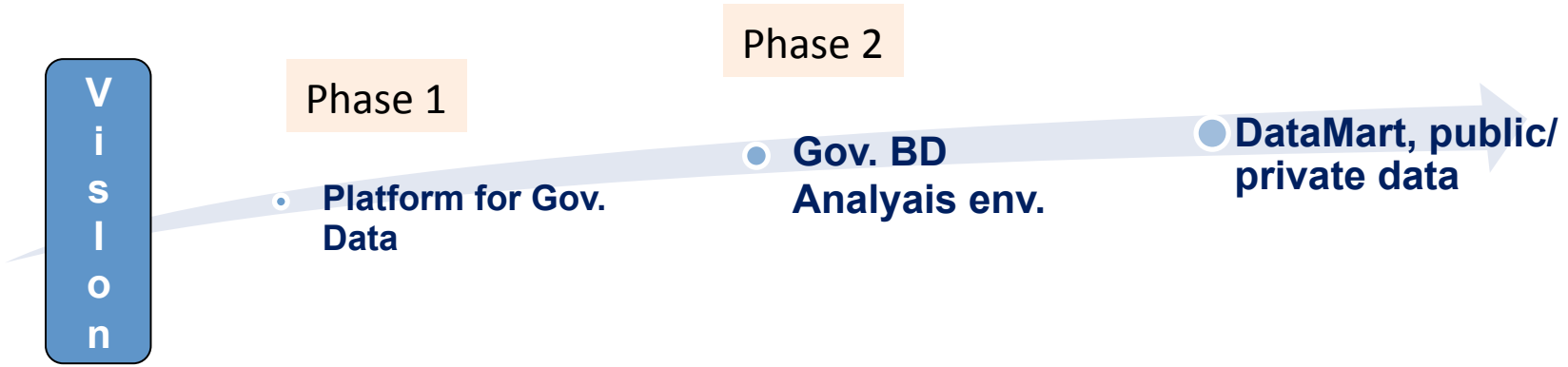
High performance platform on demand

- Shared big data analysis tools, w/updated version
- Reduced IT efforts, training time & cost

Result deplay

- Data processing mechanism is taken cared of
- Data analysis via web page

資料市集三階段發展重點

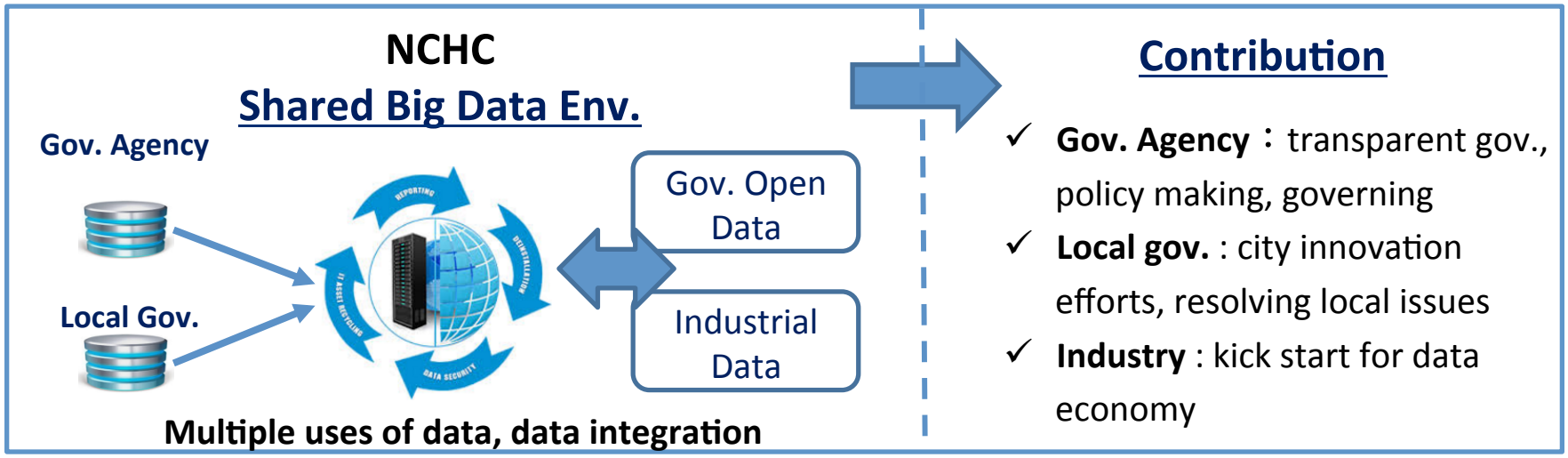


Major Work

Data Upload	Data Exchange	Data Application
• Identify data source • Data upload mechanism & legal document/contract • Data management envo.(storage, de-ID, data integration)	• Enhancement of the quality & the quantity of data uploaded • Data application via merging of various data sets. • Data sharing mech. & legal doc. • Data analysis infra., tools, env. (Data hub, data screening, …etc.)	• Data application promotion • Legal management of Data analysis results • Data analysis module & tool (vis. Analysis function)

Shared Data Env., for Gov. data application

goal		Non-open data from gov. agencies, co-located to NCHC's shared env., Provide to gov. and academia application
merit	Env.	• Secure & flexible data analysis env. & tools w/lowered hardware & software investment and data management cost • Simplified gov. data integration for better data application efficiency
	Ap.	• Identify core data sets that meet the demand from gov. agencies (Top-down research topics) & public concerns (such as education, society security, safety, … etc.) • Gov. agencies identify research topics, academia answer the call for proposal, MOST support the research projects



Conclusion – Legal Issues

- Goal：Recommendation of legal adjustment for big data application
- Based on :
  - Study of International cases
  - Study of Domestic law & regulation
- Resolution
  - de-ID
  - Case by case investigation for data from gov. agencies, No single solution that fits all
    - Purpose of data collection, nature of the data, data policy of gov. agency
  - Merging of Gov. data and civilian data
  - Single Gov. entity or single platform
  - One-stop service

Mechanism

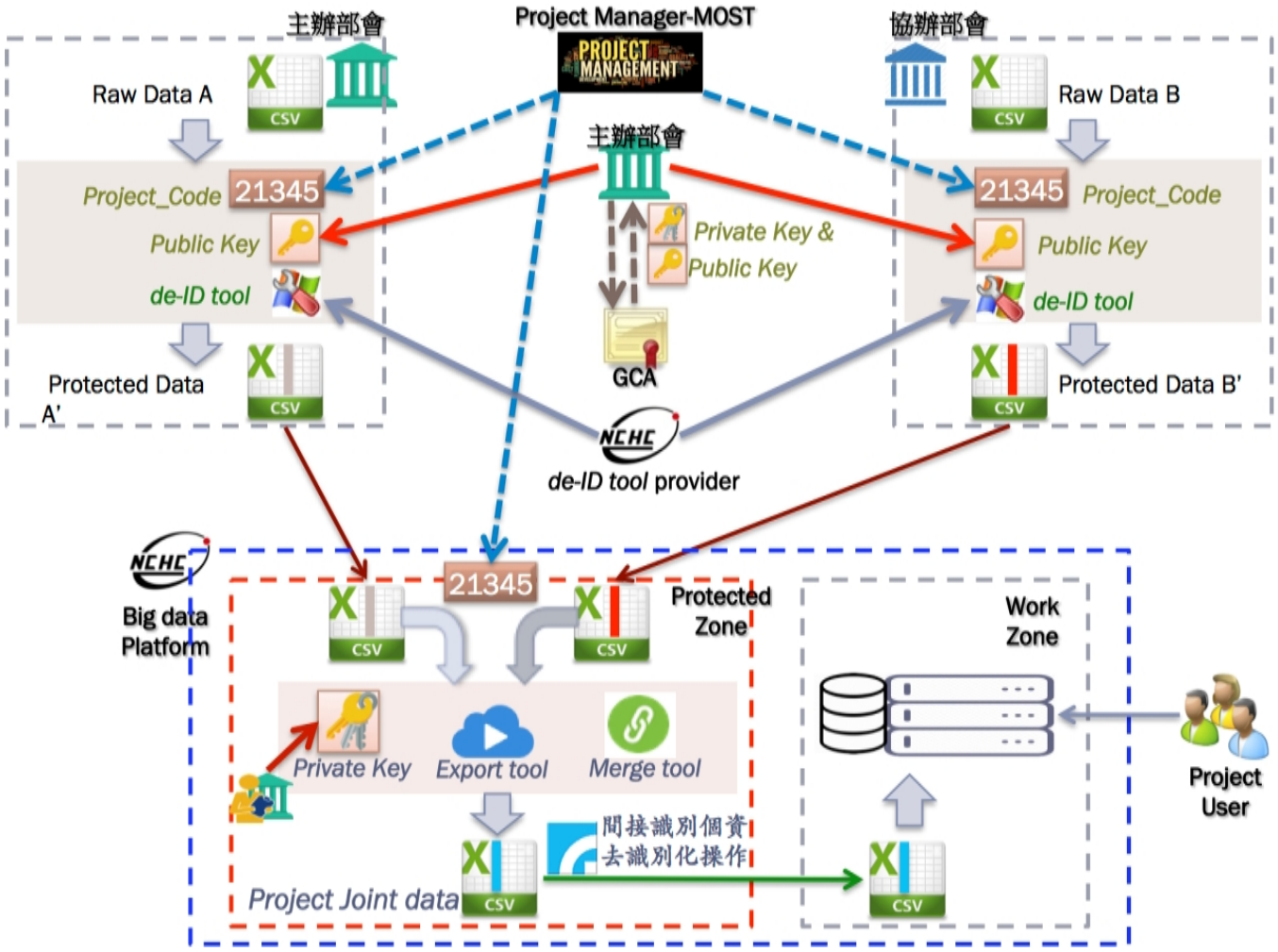
- Goal：Recommendation on the mechanism for the project
  - Surrounding the development around specific application, with 4 major arenas
    - Disaster Mitigation, Security, Traffic, Medical and Health application
  - Delivered by integrated project or data competition
- Based on
  - Study of International cases related to common data analysis environment as well as cases for data marketplace
  - Study of the status of Domestic data applications

Tools & Software

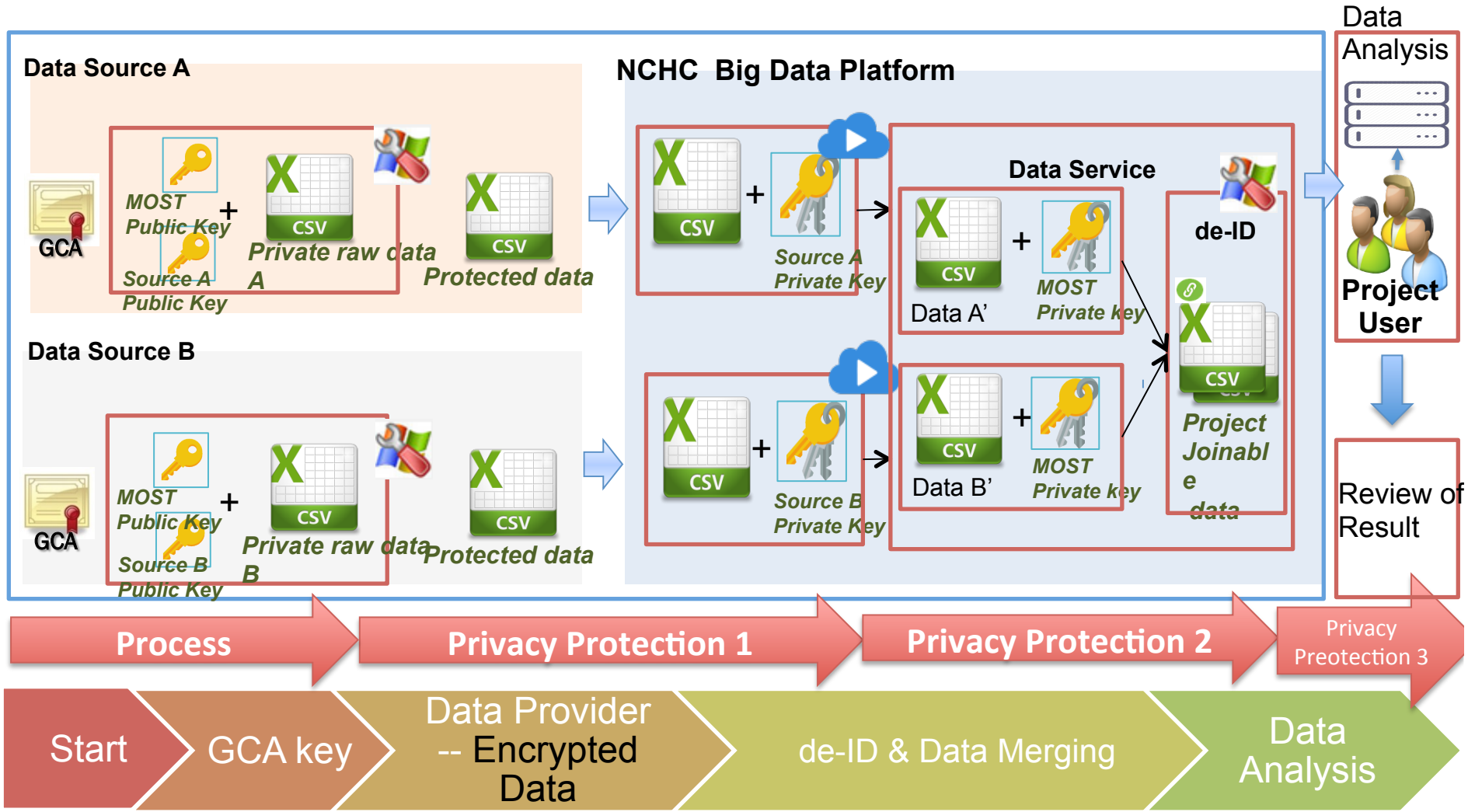
- Goal：provide sensitive data from Gov. agencies for academia to analyze
- Tools：
  - Mechanism for de-Identification
    - Randomizing of direct IDentification information
    - Anonymizing in-direct IDentification information
  - Data integration across different agencies\

Tools & Software

- Mechanism for Randomizing the direct ID



Tools & Software



Tools & Software

- Tool for Randomizing the direct ID
  - Provided to the data owner

Tools & Software

- Example of Randomized direct ID

Original Data File

身份證字號,地址
B391987589,40646臺中市北屯區 北屯路 雙 30號至 56巷
A399274755,40678臺中市北屯區 陳平一街 單 133號至 139號
G399741197,30348新竹縣湖口鄉 長富路２段 全
D399078461,10843臺北市萬華區 峨眉街 雙 120號以上
G397129977,51442彰化縣溪湖鎮 民生街 全
F399293757,26144宜蘭縣頭城鎮 武營路 全
H399225240,32060桃園市中壢區 內定二街 全
B392762833,32656桃園市楊梅區 上陰影窩 連 62號至 81之 1號
B399510515,33456桃園市八德區 高城三街 全
F393098274,41447臺中市烏日區 自強街 全
F395266496,24841新北市五股區 成泰路３段 單 579號以上
H392706057,30266新竹縣竹北市 新寮街 單 73號以上
H391994284,30642新竹縣關西鎮 新德街 全
B395922966,26546宜蘭縣羅東鎮 新興街 全
H394767850,43252臺中市大肚區 華山路 連 57號至 160號
G393113115,40644臺中市北屯區 軍榮二街 全
E399933808,33064桃園市桃園區 南豐二街 全
D398245299,40859臺中市南屯區 文山二街 全
D397441495,24890新北市新莊區 五工五路 雙 8號以下
C396056651,33372桃園市龜山區 楓樹三街 全
C391552012,50243彰化縣芬園鄉 大彰路２段 連 934號以下
A391468140,23542新北市中和區 德光路 77巷全
H398472772,32053桃園市中壢區 中正路３段 220巷連 501號以上
D397629166,42350臺中市東勢區 東關路友誼二巷 全
C391916212,35244苗栗縣三灣鄉 石馬店 全
C391581757,33849桃園市蘆竹區 蘆竹街 全
G391673147,30352新竹縣湖口鄉 大同路 全
C391532012,35059苗栗縣竹南鎮 口墘圍 全
H398687746,33862桃園市蘆竹區 大福街 全
C397015449,50862彰化縣和美鎮 愛永街 全
A391615434,11570臺北市南港區 南港路３段 雙 54號至 80號
D394984537,42076臺中市豐原區 豐東路 250巷全

de-ID Data File

deId身份證字號,地址
L1nGYbEPs2ppfigahp/03zNLvHYxOLQYwZn5y1fsSDChpy/DdXlzubKn3GpjerIQxeAeKKFSpY7SyXKO4MdWND
etEQ4m0q/1fuidYsCxSv6b7TudQq/5v7LGgH8fYFw9CMszznaTOU5kexYh6M/J5CKgUpmOxOUXBbFXILOghynN
bTKpWe2s3xv748yTN/opqKd42aSHjd5EwfVJMWVRjjusCNL1mQoJB6VJGRhb7mwT8RbzbbsfQfmB3VoCswQpc9
m9sUZuLjbo004FffnGCvoDiaryURCMnuOtoDCASP8U+TO/S4cSVBJKOSHeOnANqc4aj/+cNqCetrtVPe6ufBfA
h6LTJjnQDCmUPEgVEbSbzroj4zLNe9g8aBpaoSrpAJQ9dDuw4ivzaOAGddZ1qQFFOSFwcNVCZGoekM2/3jtVgc
Z5DOC0dyKZiKlXCifGMmEcoHkpXLsR63Ve/BRETFt+hJo/WC/zHkCfxZfkaIrOILTOLMH8efH+klJnVtARu/7f
LL+vVVdZXlqKQilxWRvzAhYBVGtZjc8WGRdzfHEvELXMZMtxzcBlRHxb6/9wG91HmzDosKOSFEnuSDndVOX0QY
LkWjJj9csVquOAI7zce4iuO54G6GP33IN+4H+DMFOkBhbblkuA+9vhD9VZ1u2cLRy+CsjKMya9rjp0Imf+vWHY
O3DPQguGp1Nz06dM6zYGS2vmydijnc/iILJJzXLYuM2K3mayzqcR/eGbAUQ1zTkVU/v8XUoByOuzqsXJPNY5Oh
JLGZw5900JAQgyOLK+gMR5cbVPnE66jNmqBO9LDNURXmsBSNanHdbivlx3ZOAhiEzfRDvWREB8aLSdLYeRXfvc
eG8ofVBbRIw75umKmnngPBlec68Bg4s02zKx3jmtRX7Jh1uCa3T5g6+T2960wNRPOA8Jo2RZ+jpjbgDulLJaL8
jaBDoCvtmxHATthu34bLnCzxKlfOVY++yn2YiPMOuNMIiCPBzUGzdjntLCvlufxlr7qtFc1NRQUCcd2U4Ft8H/
WHK/BUTmSF2/UM/w7IsPQdYQ1ud9SjW5bNyuk6yKs3rb5JqFe9LqnkwNJ49tdInN+HzstWwIzZJRNdEciFn6DB
d+KbhrV2n60j+4l+DkLrGZ6oF44LKlEcSrOpIXJQ2y8DpNrPCNeDKxC8ApYHj1M2gsSXY6ZeVmQsVIIex/mYAh
WSEhucy2YCia+dWa4i4vlDfen2y3s+IG0Aw7nM1Rb8oMrMo10d1zK+gRFnRc0491Jy2FiMWv6SH169VxTKqKV9
XuCruUIPWOXtrficdgBAJlMp02u8xS+4cMJJsbZi78mPjxx+H1VWn8XfHsVbVHHDMMfL3cKLRcnoVLmzXhkFGy
R4AIZZ5QOseINfVONegdAmrGWoIPN/UR4tZYxzTfgM5Xn6zSNWzRG1uk9MgjDb9QU87YdAQ6DYnPScn1Otlm19
FicewsBALbgPeXOB3hFGuzN/6tH9mUNIarWbjq5IBAXBHfoSxeiSJJMujANE9uyHgtuNeLamNh6g/uTOZIEgB6
bUZKwkAmGsEQZff7XYXG1uZOkBNicwHjaWUBHYiwZDOW2FSDw2raM4PN0iYQFELyviuEQVg6h6XkglGt1GDpLM
c0R/G6g+lbxltoDqJiNQuKBN6znLjBdZ+Pxkgqwz Fy/+4mlem8pSOrlbO+tq2HsuvzL9ZtqV3KT44KAckhY2tS
hlm1c51OArJLhVR58OkHJ4R5gUTTu7CF3XPtRV+vRGQLVQ/vHQ225+6sfuE5YLDgYjATr643QOeqiyy0BmTZ39
LpHU7bYzqFwd2mqYRBTkXbiG5suww7f/5OhBZ6WfPmTlutJ4Zyxfn+hGDx+sKMU19Y1Rr2UcEK3SyQlDrndhhw
SL9FzgNUzySfmCygsmkupRoe7y+4zBiX7bsM+qGMXVDtb8KaEbaSLyXhV3C1IORbTuQ95190CxQno8qTMKVMEu
L6RI0ekIMv9UrQEoqroFT193CV2QEG1LcmfC0soxtQxbIE4jCwrfEkd1EHLCevGzPR0VtGjw2tGYKeStIjAIfv
LnleGiGNEARtLtaec7ae4cCYMaXet/4zELYY98HAHQo4K2ZlWpTZVbzanS7O2gQDoTzCs3fcB38pH/wyBVULLH
SfRNzZXwooDlVL7jhCFAzAPlzkiuNhp2JgKTZucwsCIcmaxD/Z///4PxluCIWiUSfiOTMsLBDfBwzNBvs6OtqN
m4ztprXDxUTYvrqptRAzUqSb03Ih7JYCieZORCFlhRTBf7mf7yYpDEnL3KZLG2p6zltz+D0qRbjzQCIlUDTI/W
B3kLQ6o2Mhb2po8ix8AC0IC1UaIKF3yNIZi/7C4g8/ITb2g3ROzwcjBKeWgljTek/MrKTpxwBiVxxbObCVwI7N
b+rTEDRKr6LAtvDHDobYODgOQFSzgbyP7dK9Tx6RhXcQ+l7J55qmewCPdMBs6ttgVCbCobSYbpFwL7Htkh4Caf
eEaxnKP15+aNWNDbreN8chPqsPb2bxOCaxF7vJMVUajjU36nBh2HKKuhBUbSw9VdIbwso+Kal/SQ7pneksIQkM
EIrtwZzDOPJlGiljfJl9Xo2N/SCp2+Zoh5NZdS9/vzGwGBojeB6FEoODz7oNRT5ZE2l6NEX5OQCxP9Ox7MaBGH
KWkEkWaPp9g+uufEY4N8f2HbKkpUQq9CYdTkjG6+5QVKBVqY5iX2Z5RHCcqnw9OhWnRz2EQ8wiHe6hWc6AI5dP

Tools & Software

- Example of Randomized direct ID
  - the rainbow table

de-ID & encrypted ID	Original ID
9zH15giGTFtjm81LcgHFVHTrXWYE4JyoRGb9YXE3tJNiOzeXGP8U5hMcnOjal6+ke0A5f2BvjGGis=	B391987589
mga+sGfMDXCjKEP1Emcd1duPXfvI+3XTLpds6uKPNL3i6yXhOZUOrEx6y+vLRtRNlNxXlAzAwKbjw=	A399274755
CRGWsEximwKnS98YLgc0LjKSvpukoPJ2vYANcZ/gi1uRXHEA/Cy96H5omBHtTynR0R5pL6n4riAwY=	G399741197
+ZkgaOdVA75i+u6UNF/0AaEiG+maFoHP1Y6kqWRaewxlQQWVlCV4DqA9lz4+PTUXMsA0gKwyftdzI=	D399078461
1E3yq7vbFMUl+/Nrppm6+YfA0icnFAI2zZyIWKUa254Pg2pmeeqahXPfxRX5XER7iMeVNi2XzbWII=	G397129977
PY6Gha3+mLtWYBuPitP+w6isGS19rtS/W+uKYBHiAz3sDg5s1QV2vBQXU+8yOLk4sTO4Fono6C6IA=	F399293757
WrIZAg+KiwCuXNDQIf0iaHWxIku5DcuFxMtpgoD+CjplG0e6T6sIJZIjXBw7E0IZovr7QL5Z/J8UY=	H399225240
TAYxA7ImoOzZWlBwNxcNd8/eE7MGVU9ZRS+is17dTLTj78gQA/Mxgy0/VC60CgvKYIkO5mh8n/4tI=	B392762833
i9VDG284SqtdA5JGY53G6b1puEkEial8go4Q0bKlBGw7e02F7x6TqZeE1eoMWFwHEPoZ4cKVk0Ihg=	B399510515
le1h/8EE8Q/Dqp26siWrpl7iG6jSeCI+FOkz9rDl1m3fs02pab5bKDeQYswHXcX21VWzNxvvj1Db0=	F393098274
NmuYW6k/sd9hCXwHI5ZIC5okN54tp0QbbXNYJ7c16McIZPBM59Tbmikvw6fxzgY1eTAo5iONwmLhk=	F395266496
9XXb+jqzqpEYgiZUjOoNz1Bbrt66A4SGgBKrsAYSbnbRTPG0KikTC+qeTjxTERXOsCukGjKoWwO3w=	H392706057
RQZn6M15DIvgcy6tWX4OI2mNZ+HrEf6T79Z3afZIOneCMBTIU8eH09MrWarhUfMfoKUboBx0htUjk=	H391994284
A9dmhrYb0E2DNiQJLHRzSVxzuihSOCh07+yoYgHvwGEP+6atUFIEkJMq72uCKM7pnZh4qNpc9kSt0=	B395922966
kygpd4MBspF1OshrOunsLvQ1goMbxLihJ6SKkY2OhyfRnzAODLCCQJqpqR37PG77k3+PJBS9Yzdis=	H394767850
/6XtDnujKHj9e1ZG2B9bU3SOhmjB6oa+t0ZII+ephBuecjtvCj0XEIGXbeFjUyYbwtlWfqbLbAcwO=	G393113115
auGfczcCNVR7aMzaeHsg4eOMrxN22q1QtUzJHyNzivXbHXYLxLqE1MwEPDbyN1BOPBDVY7xLLCROg=	E399933808
J7MaZpNBNjPF3VU9Ky/WyewxwvUswHMkTNTSO19nCdJvbERiCC6VX41ZyhKTsx/HXO184eo+BJXtQ=	D398245299
FvKzUNrcdRUiDnZEfzvBZCZ1QVVrQ9CeBy2+mMNYpoe/4lqm4bwq2g1N8rq2n3JgWmlBZrHC769j8=	D397441495
bwEIWkadPfVb+P2jiIYiyLnRpZT5QCdtCFxUVgY5D/kOVRljIZXBLVDj1BQlqzjYlxYUpnExD8V+I=	C396056651
3PWZD+3TBGLgYyrsLQNGgezhOuXqOEUk5XQaOMQPPsaPuK+HabJY8iyOCg6SR7pXbF1C8h9fpGaoo=	C391552012
fOM2TrWC1MbbantltPoGZXn9eERcT16hyBOWm8/agPO3VOLpCgwSbbxmuc+J4yiaba2+HWFizWd24=	A391468140
EjQZjmRsUs6uPFmzhBWwRpKOM2c15cHaFivnW+su/TapEtQgorG5TixUvp9vVBQlzVozs5u+XSkgY=	H398472772
Dyfu124Qw0SzZRYCvLM6yt5X9XYn7cp/ofwnkySNTEtTKY4NAsoiBkDldYwA36pe1mBOqad9bjLAg=	D397629166
FouoqUVwaPksRO9ugVtA5UlzDreuv8b50w8/9LPzoQJiG8IUB+b+P7JLBMVyi/AcjmloqbMXWR1HE=	C391916212
5E01f4Epng9JxQuyG8yyi+0GRKt9XsEjXOFCQ/te8DgIig0Tu4f4hwezfStToOVWJPjyzwG15a8og=	C391581757
PfVQ0qSGFvMqJ6wqa2TgtYd5jXDxTcsFMOSXlxpOZyH4ZtBndHB93O5aeAa2Red6LO1Y1iAf6IXV4=	G391673147
lNx6js1RKMqHPax8GPchgO0aQERzRg9ru4RsrE12fdhNqKq7hdsj93TPu6T1VEWtqnxwykhmqGNW4=	C391532012
8ZABkgpYgBOUnRRLeJFErfY4KHPyoawYgkJaQEsulKIOnXg7mO0LIOR/5yZWcEmfbxZ1P628KnRSO=	H398687746
8Wm7mpj28URoRoVzc8O/OPhRtCB6VO91lGBMh439aUHoBTu2xJsozVFcFRqF/whAHA/j2a78lGKow=	C397015449
9d5CoP6f2EFObol4u1fb++05ahNDiWgol6Cb4Hl/MYXwnLJ1IOLJZzX9dH3kMPnROpo5Tg26n/Ges=	A391615434
QzJMeEXdXE33ahR7m9ex7WvVooVpoJZ2IlDTbIo9KbVeFLOfzvN62I/iEXy8ZTfUJ2BCzXugVAIXE=	D394984537
6vavG4CeI31Zo35xZj2txcUn9XTbcLVFuQTG4mrRY/W22OFr3Sys1kZ5FAQsM4W/374CXyCxdh3pc=	B395293120

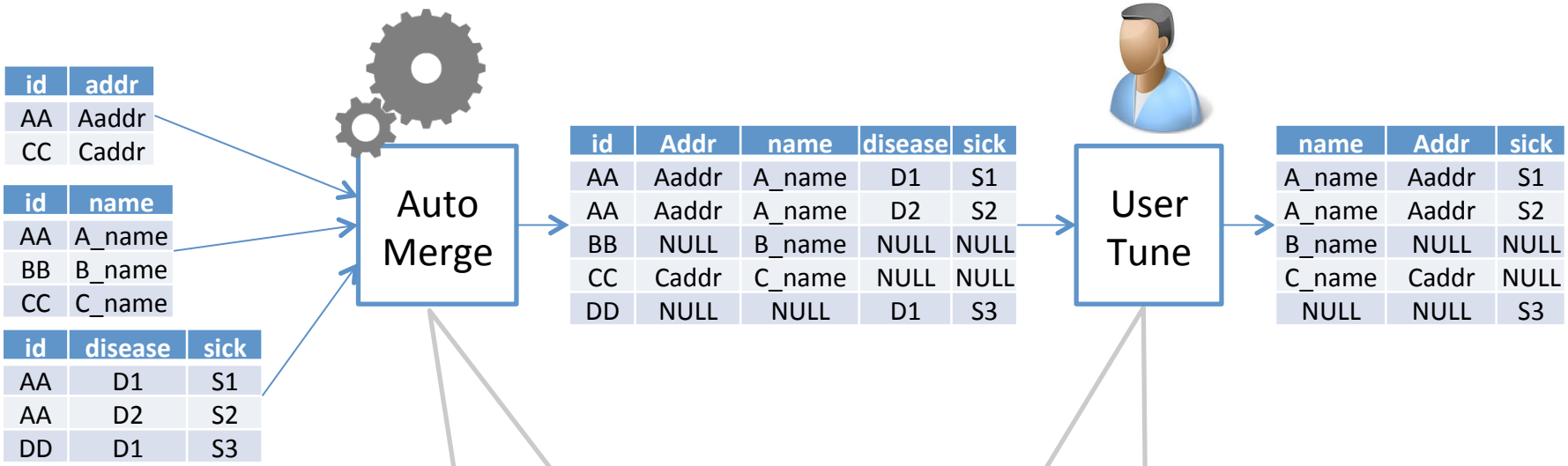
Tools & Software

- Performance of Randomizing the direct ID
  - Parallelization
    - Multi core ：processed by multi-servers simultaneously
    - MapReduce ：parallelization based on Hadoop cluster
  - Performance
    - 24 millions records
    - Non-Parallel : > 100 hrs
    - Multi-core (4 nodes) : 35hrs
    - MapReduce (13 nodes) : 17.25 hrs

Tools & Software

- Merging of data from various Gov. agencies

Design by ELT (Extract – Load – Transform)

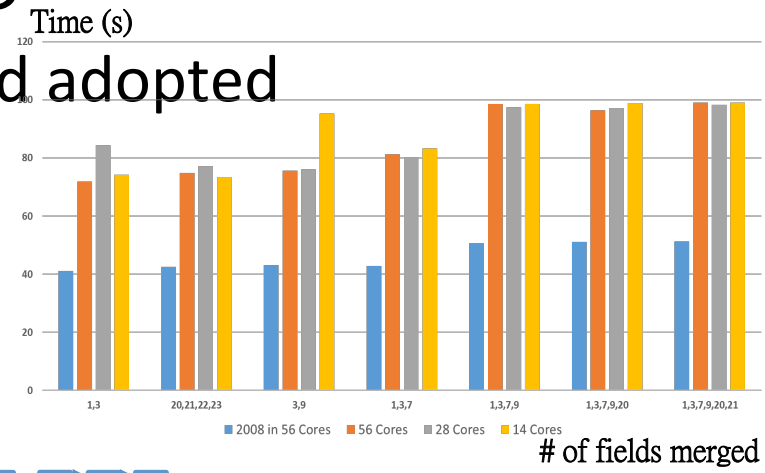


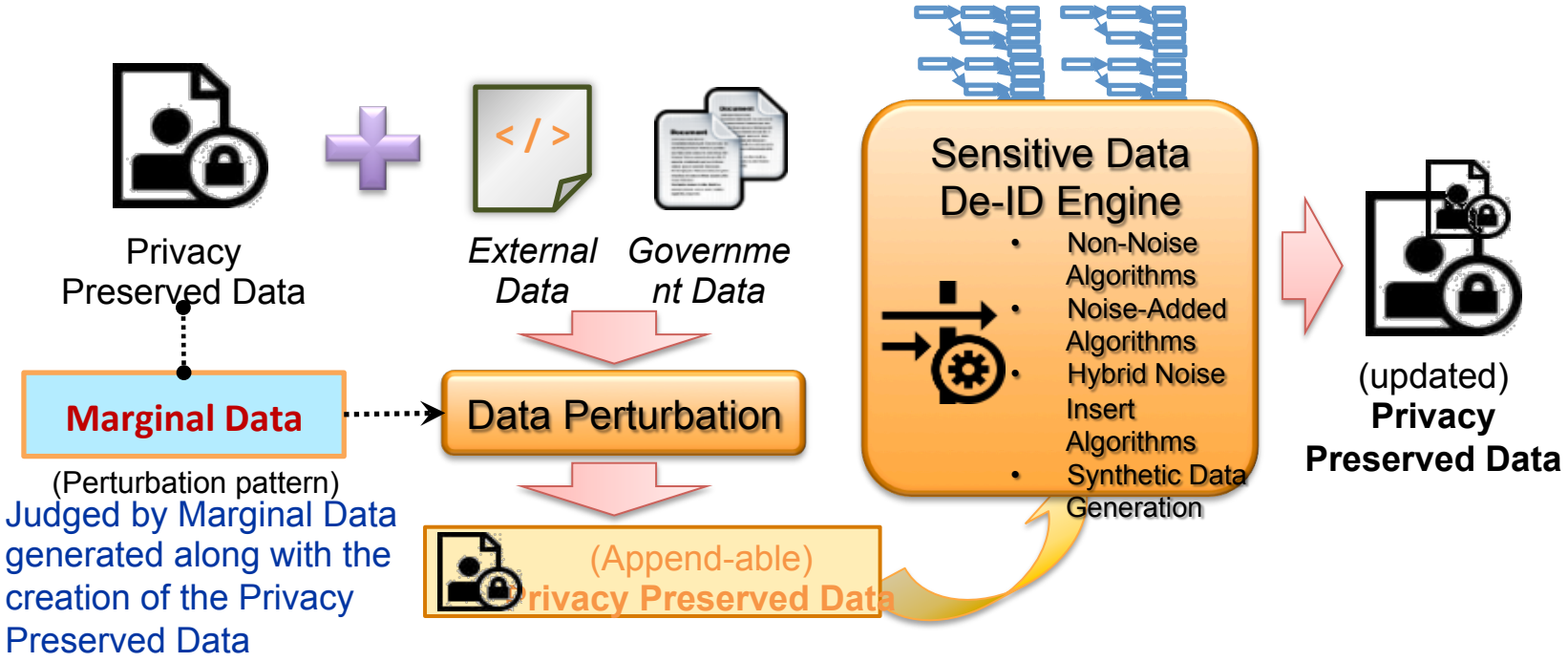
- Program to auto parse
  - # of data files
  - Fields of individual data file
  - ID field
  - Handling the field divider such as “,”
  - Maximzing the filed
  - Corresponding 1st filed

- **User need tune on web**
  - Filter fields
  - Adjust field order
  - Extra customized function should seft define the hive QL

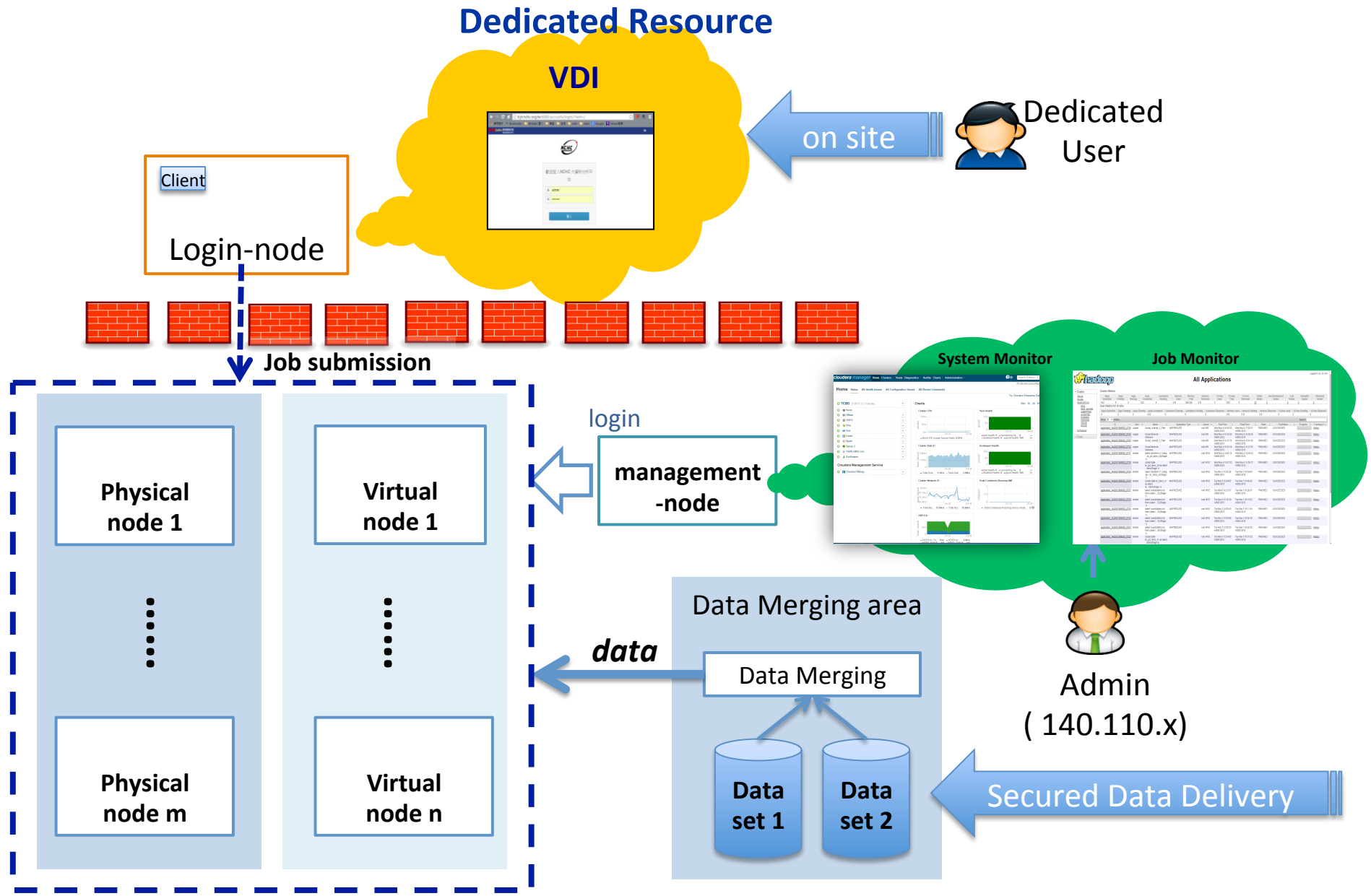
Tools & Software

- de-IDentification of in-direct ID
  - Parallelized K-anonymity method adopted
- Performance
  - 1.4 billions records in 4 ~ 5 hr
    - w/5 fields managed and verified



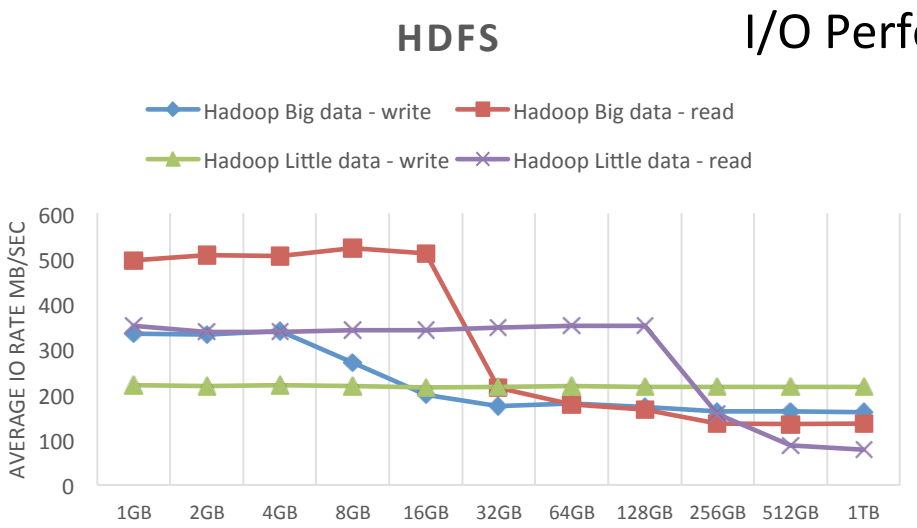


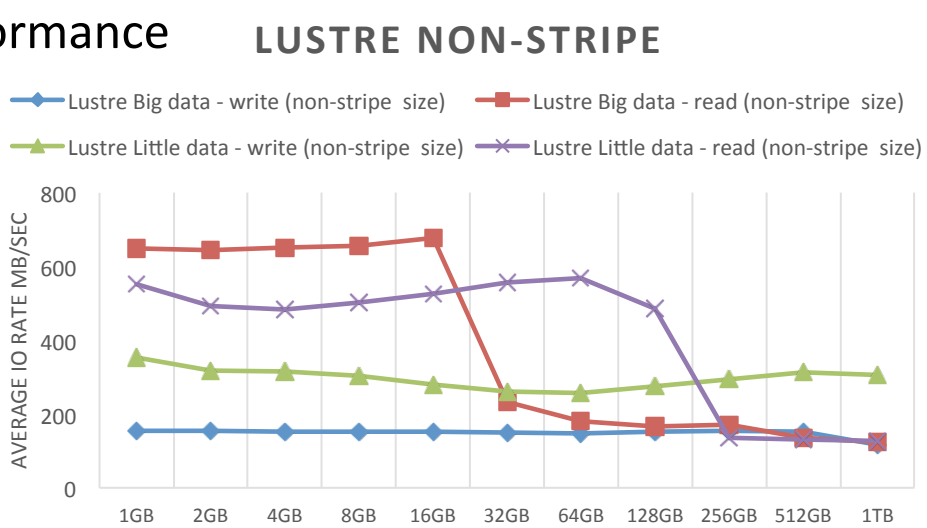
Platform & Service

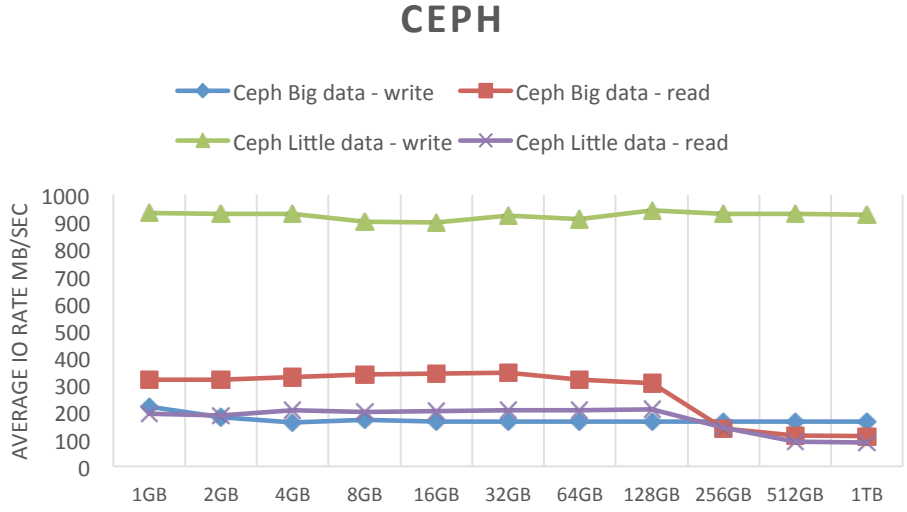


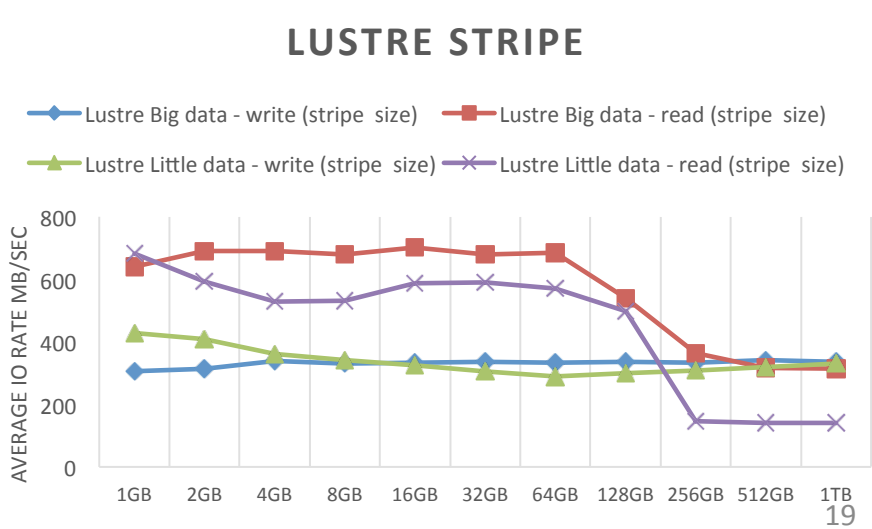
Performance Comparison of File System

I/O Performance

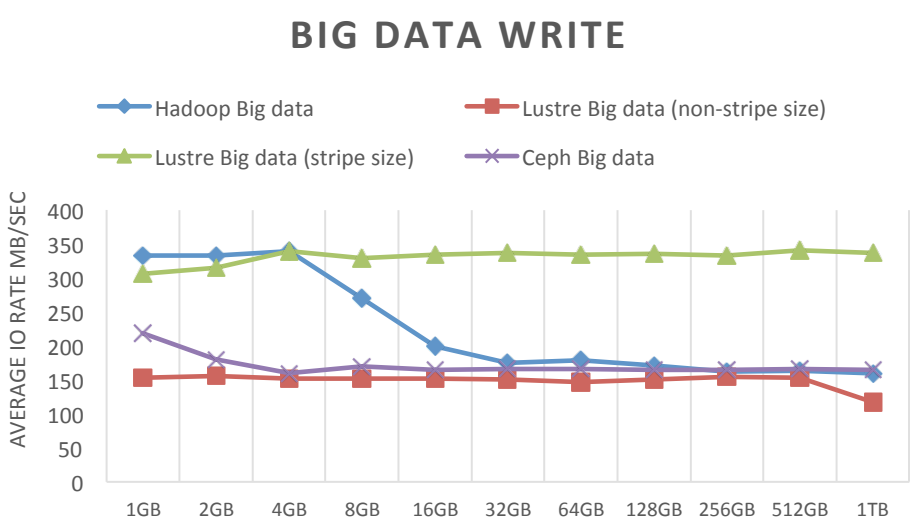


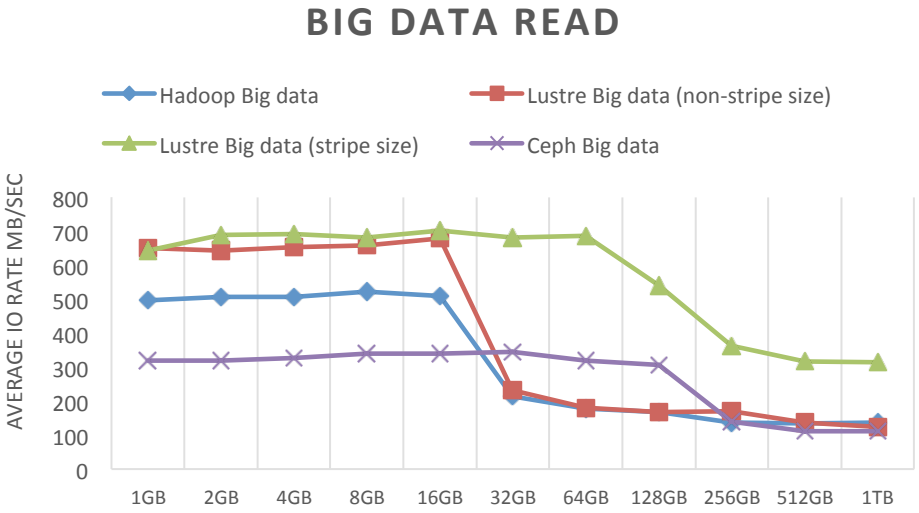


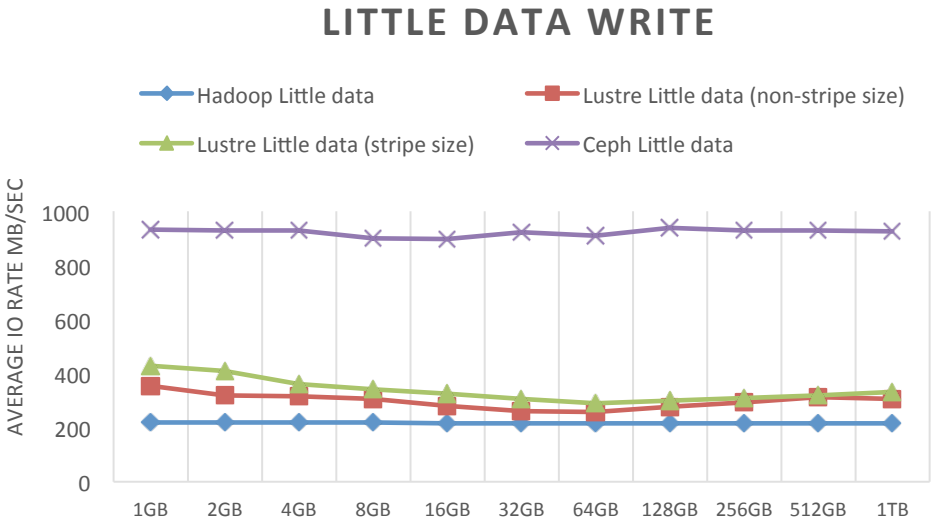


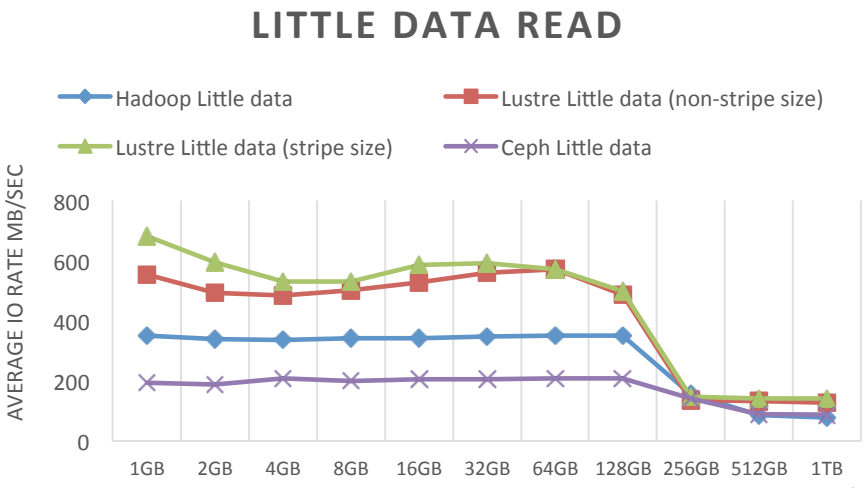


Performance Comparison of File System



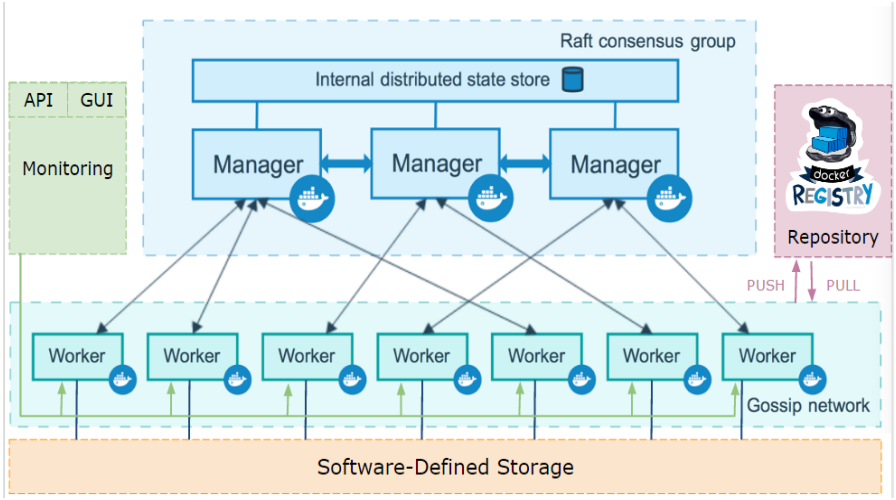


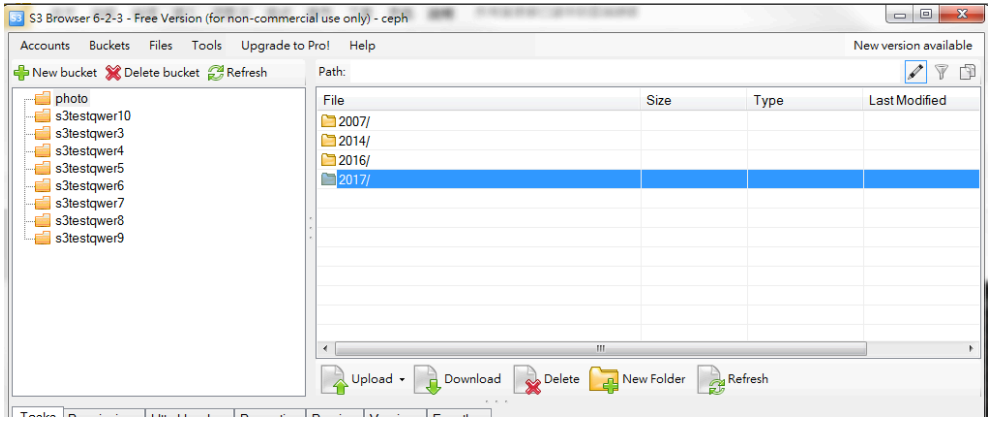


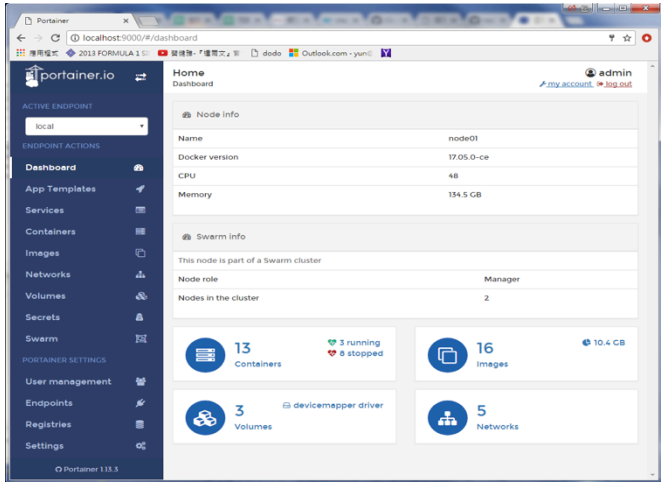


Platform & Service

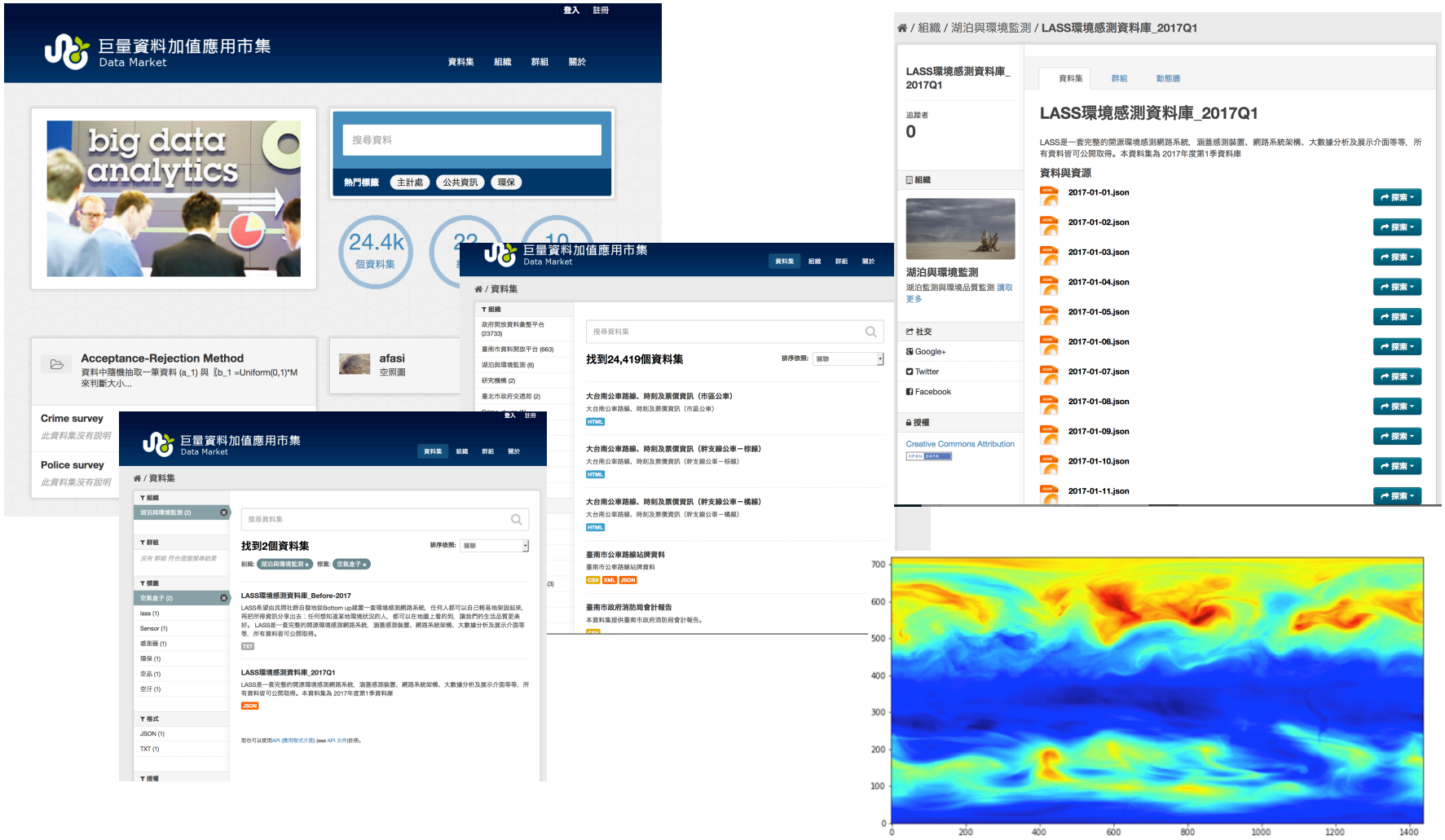
- Application Repository
  - Virtualized Data Analysis Applkication Sharing
  - Container based
    - Swarm as Manager
    - Jupyter for Interface
    - Example：GATE image
  - Ceph File System
    - Block Device for DataMart
    - Object Storage：DataMart Intererfaing







Data Marketplace Initiative

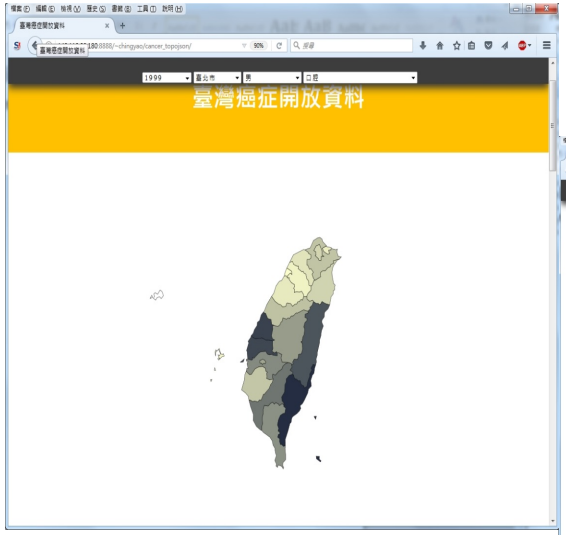


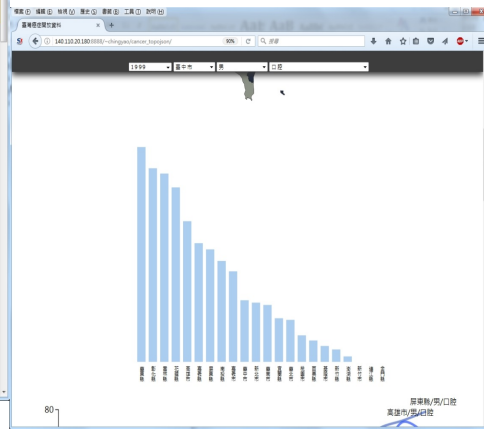
Most data : public available

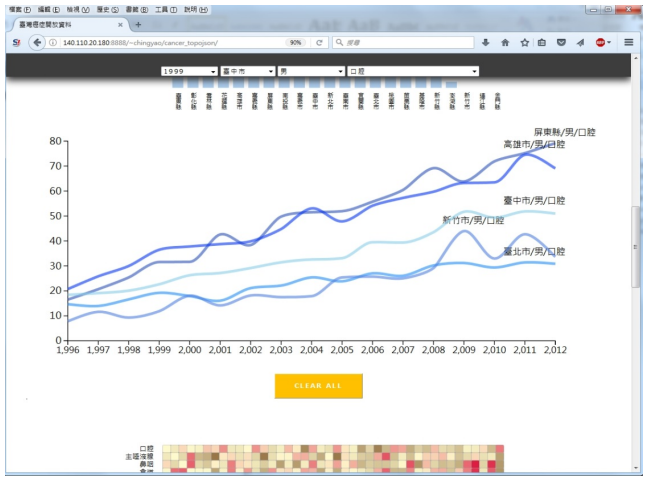
Restricted sensitive data : ~ 5TB, 5 billions records

Data Application Example

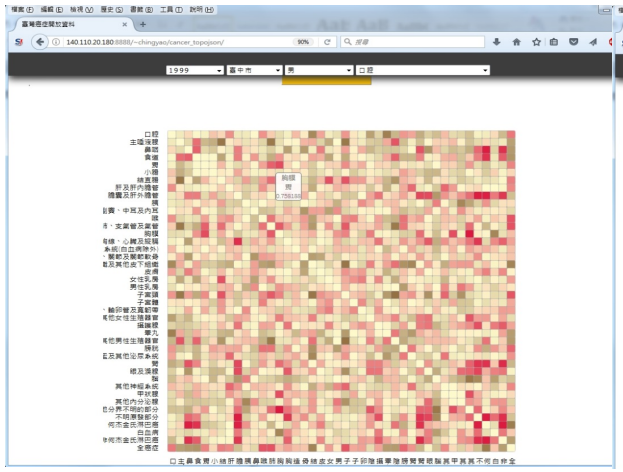
- Cancer Open Data
  - 1996 ~ 2012
  - Pierson Analysis



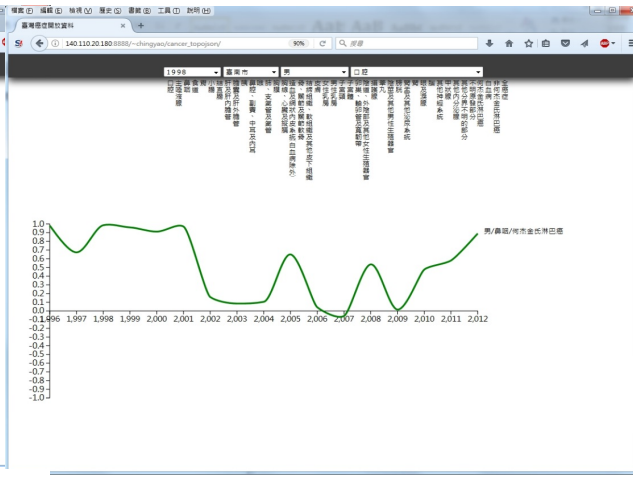




Geographical Distribution of Cancer



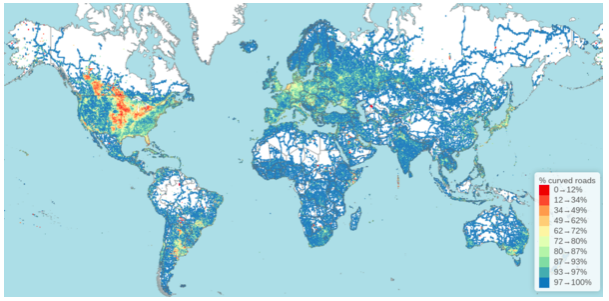
Pierson Analysis : relation between cancers



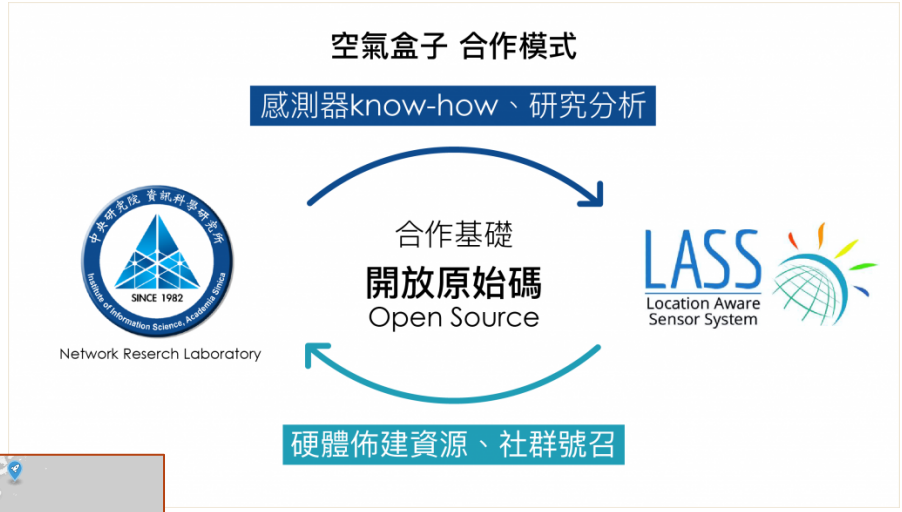
Pierson Analysis : 1995 - 2012

Data Application Example

- OpenStreetMap(OSM)
  - The only mirror site in Asia
  - Integrated with Overpass turbo, Umap, Planet
- Location Aware Sensor System (LASS)
  - AirBox for PM2.5



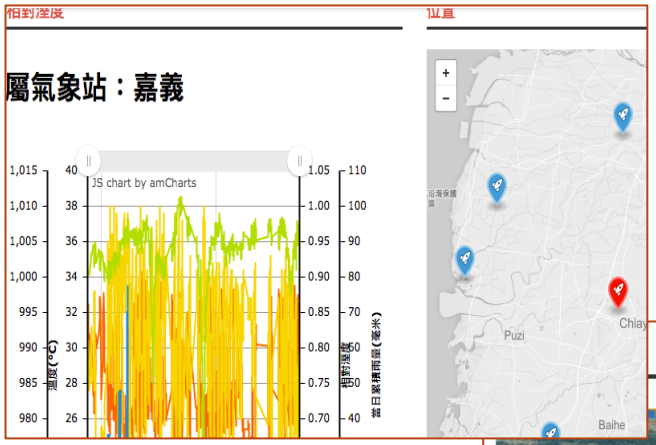
* Image source: http://www.openstreetmap.org

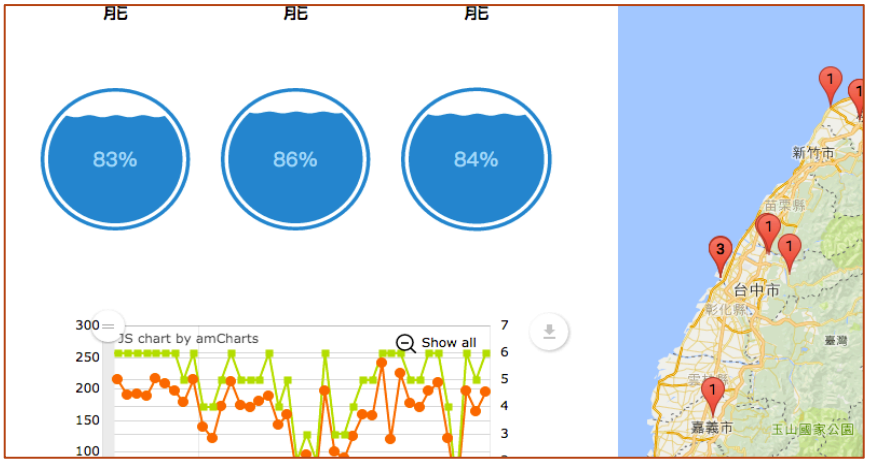


: http://research.sinica.edu.tw/pm25-air-box/

Data Application Example

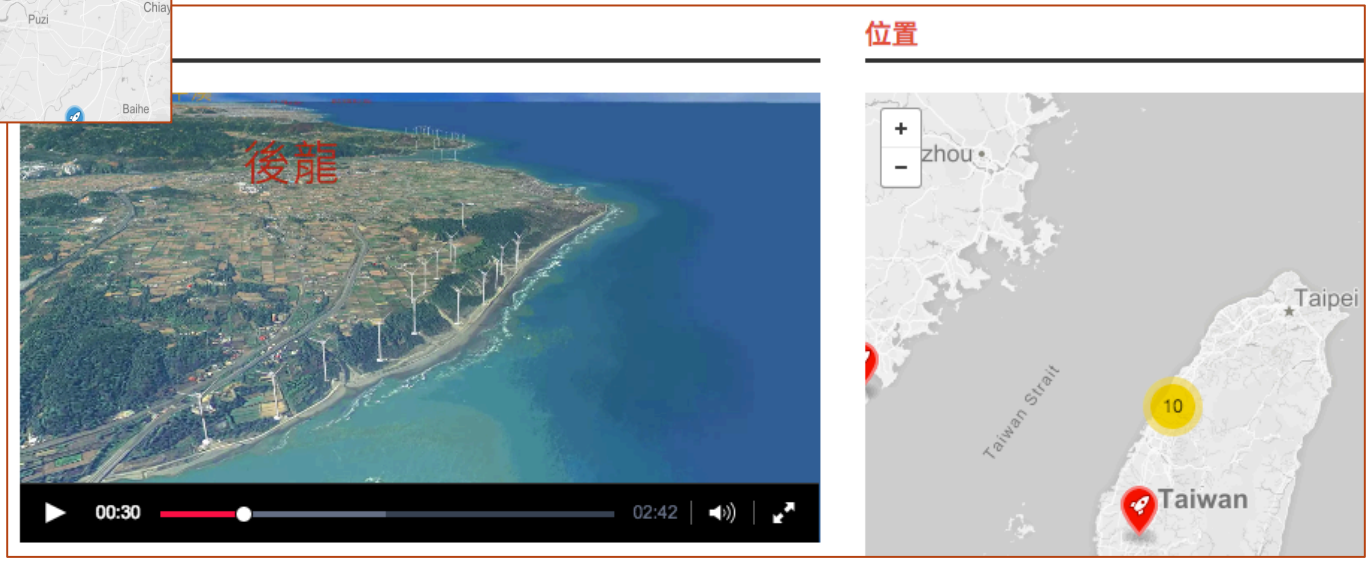
- Renewable Energy
  - Solar & Wind





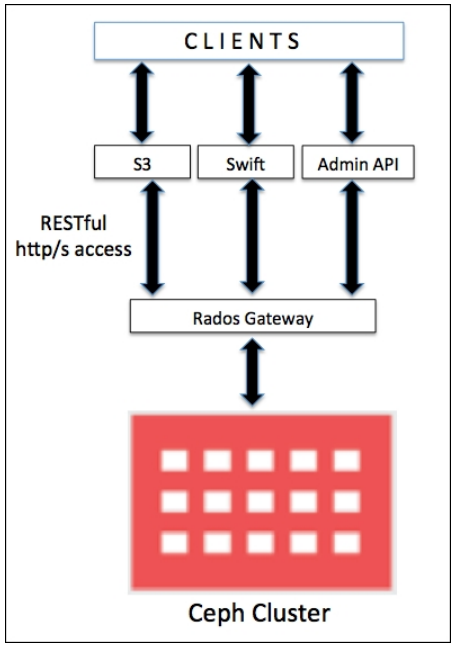
Solar energy

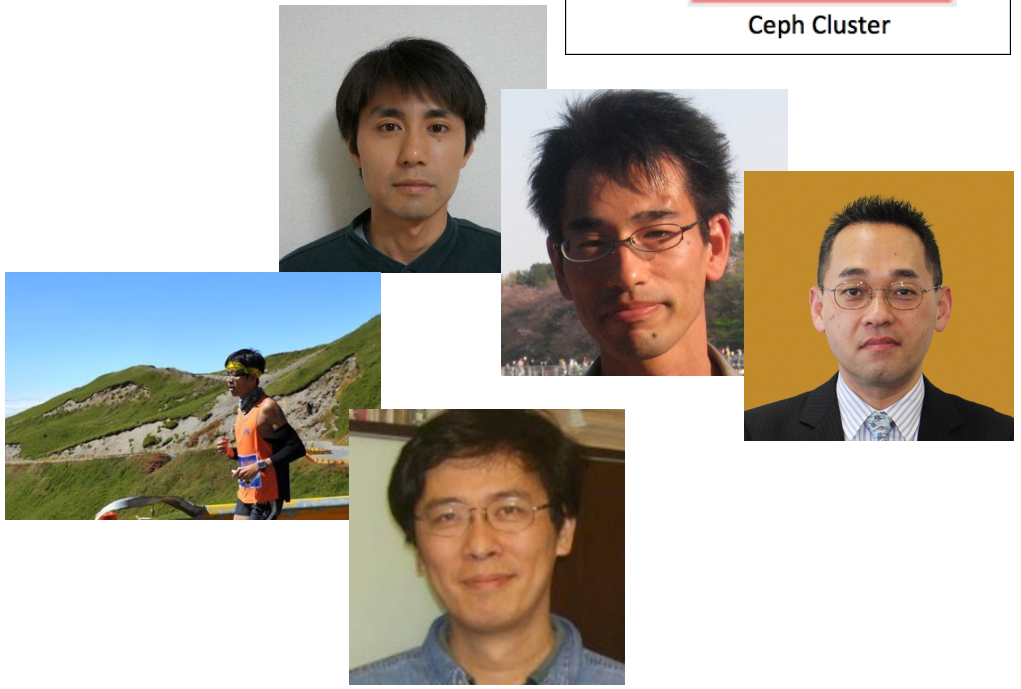
Wind energy



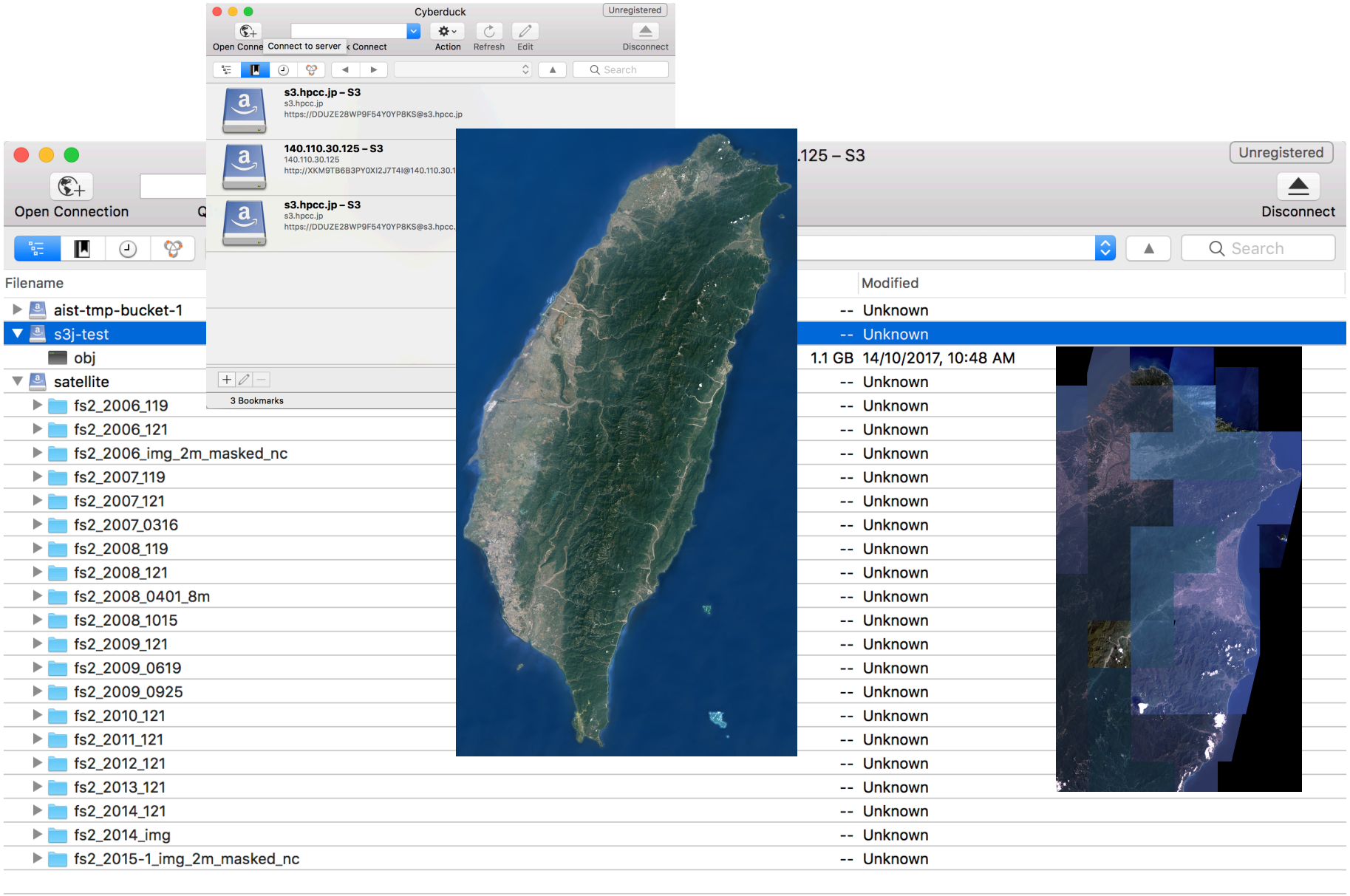
Out Reach

- Data Cloud initiative
  - International Collaboration
    - AIST Japan
      - Yusuke Tanimura
      - Ryousei Takano
      - Yoshio Tanaka
    - NCHC Taiwan
      - Max Yu
      - Weicheng Huang
  - Ceph, S3



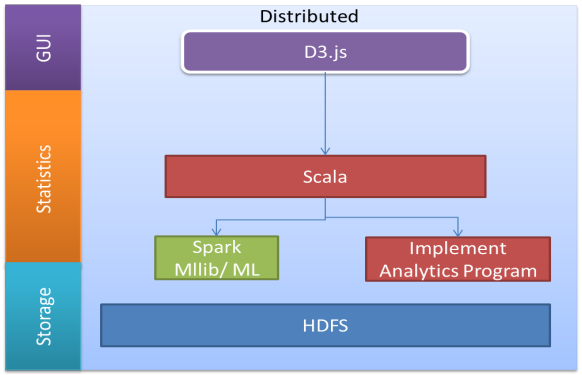


Out Reach

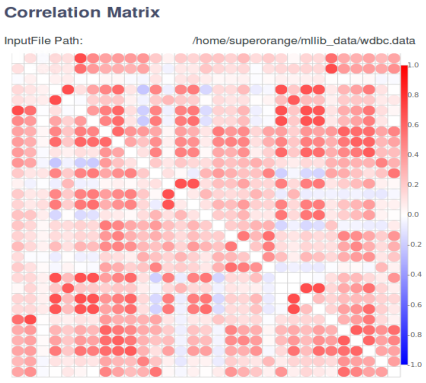

Cyberduck
Unregistered
Open Conne
Connect to server
Connect
Action
Refresh
Edit
Disconnect
Search
s3.hpcc.jp – S3
s3.hpcc.jp
https://DDUZE28WP9F54Y0YP8KS@s3.hpcc.jp
140.110.30.125 – S3
140.110.30.125
s3.hpcc.jp – S3
s3.hpcc.jp
3 Bookmarks
Open Connection
Unregistered
Disconnect
Search
Filename
Modified
aist-tmp-bucket-1
-- Unknown
s3j-test
-- Unknown
obj
1.1 GB 14/10/2017, 10:48 AM
satellite
-- Unknown
fs2_2006_119
fs2_2006_121
fs2_2006_img_2m_masked_nc
fs2_2007_119
fs2_2007_121
fs2_2007_0316
fs2_2008_119
fs2_2008_121
fs2_2008_0401_8m
fs2_2008_1015
fs2_2009_121
fs2_2009_0619
fs2_2009_0925
fs2_2010_121
fs2_2011_121
fs2_2012_121
fs2_2013_121
fs2_2014_121
fs2_2014_img
fs2_2015-1_img_2m_masked_nc

Spark for Machine Learning

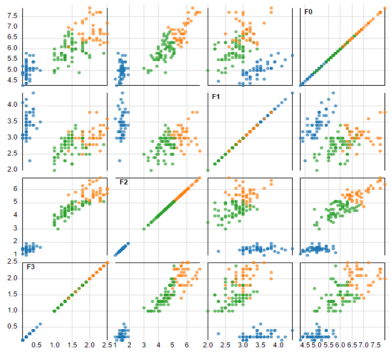
- Data Preprocessing, Analysis w/UI
  - Analysis capability via Open Source
  - ML process w/parallel processing
  - Features
    - Data Visualization : 15
    - Data Pre processing : 5
    - ML model : 4
  - To lower the barrier for using Spark
  - Better performance



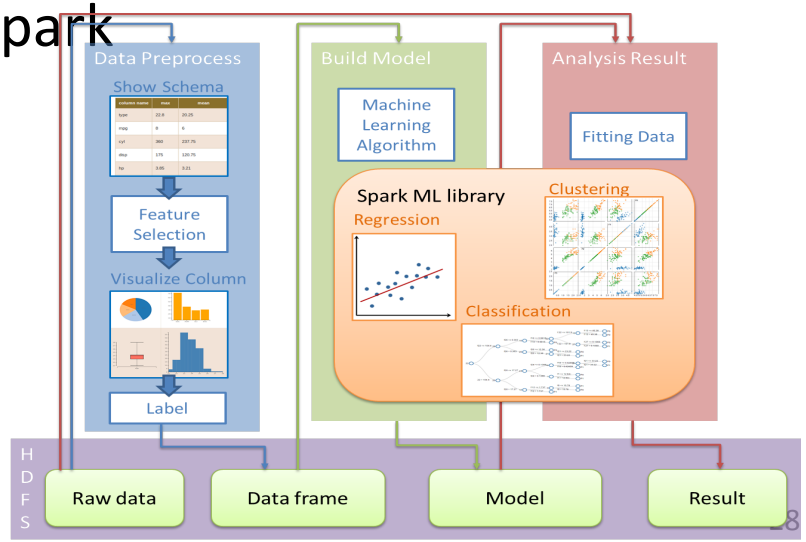
Spark ML Platform



Correlation Matrix map

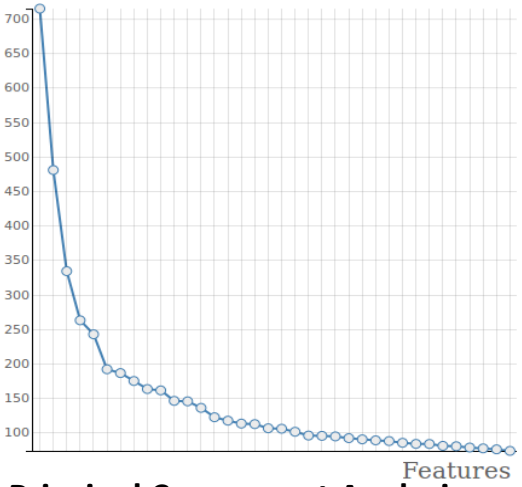


K-means

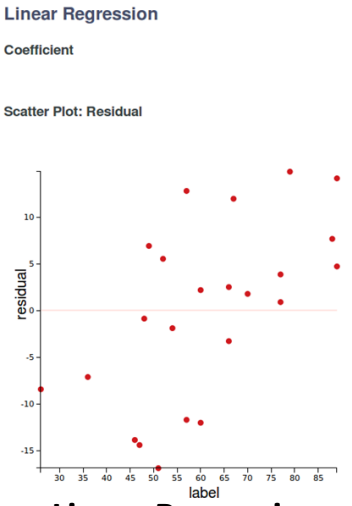


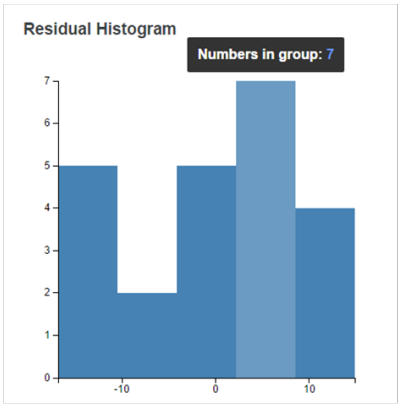
Spark ML Features

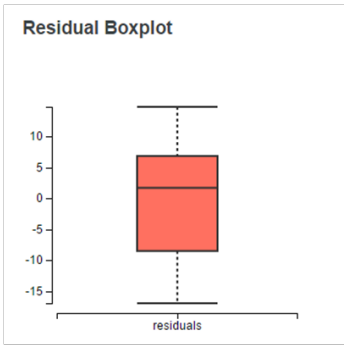
Spark – ML w/Visualization



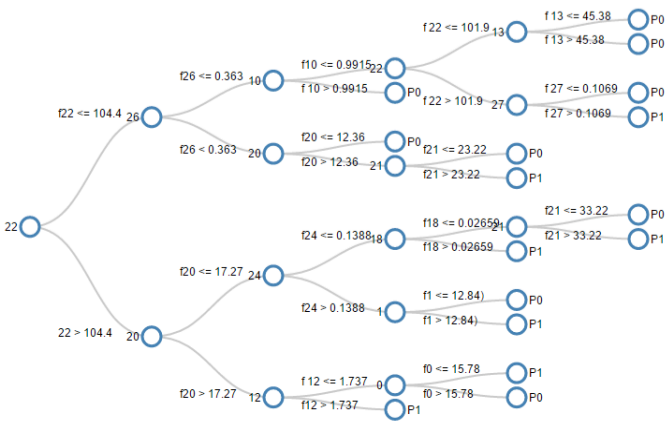
Principal Component Analysis



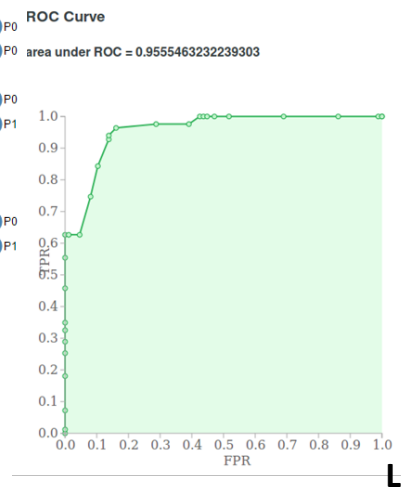


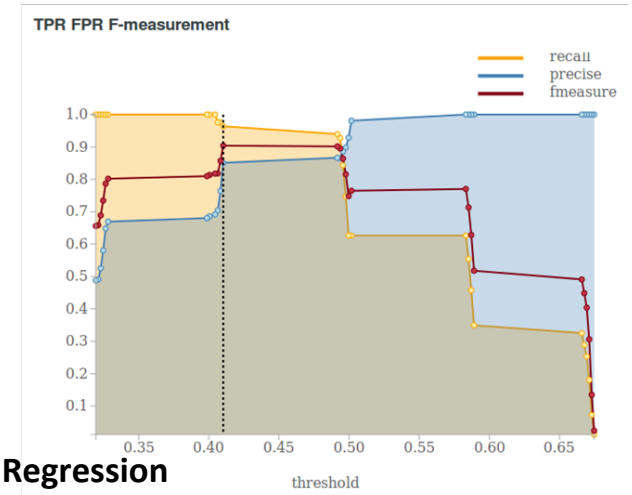


Linear Regression: residual & label, residual histogram, residual boxplot



Decision Tree

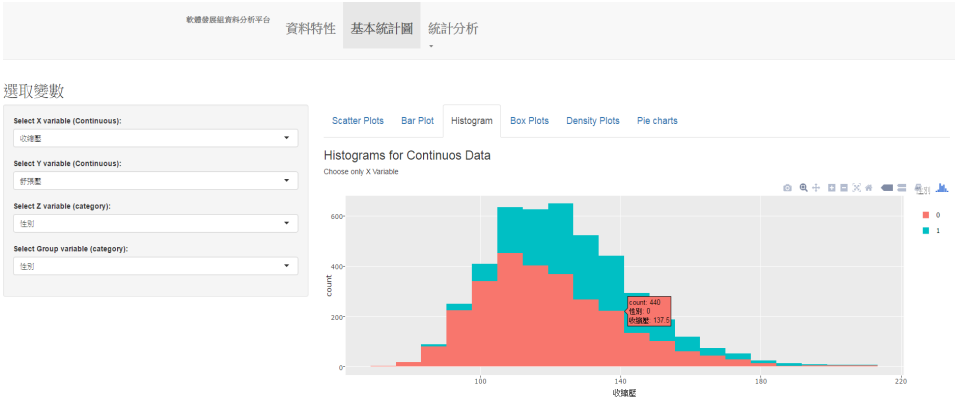




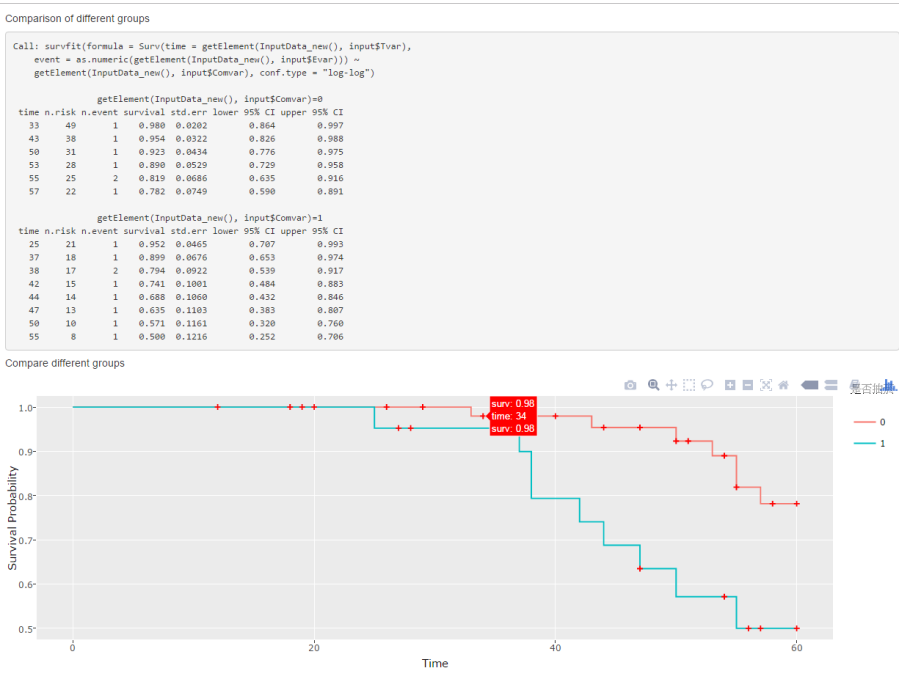
Logistic Regression

Web-based Data Analysis Platform

統計散佈圖

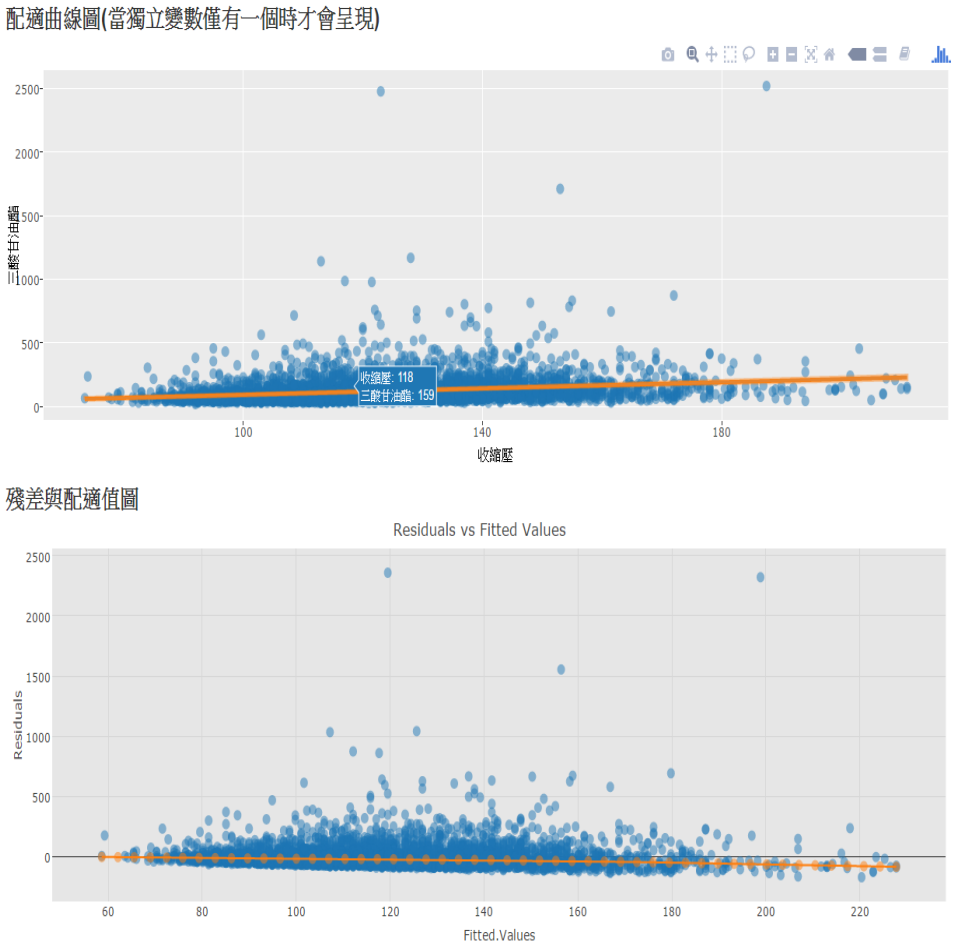


直方圖



存活分析

Web-based Data Analysis Platform



迴歸分析

Web-based Data Analysis Platform

- Data Cleaning : 3
- Data Integration
- Statistics tool kit : 8
- Statistics chats : 5
- Text mining : 4
- API w/R & Python