Cloud-hosted data analysis: an approach to hands-on training for on-line courses

Beth Plale

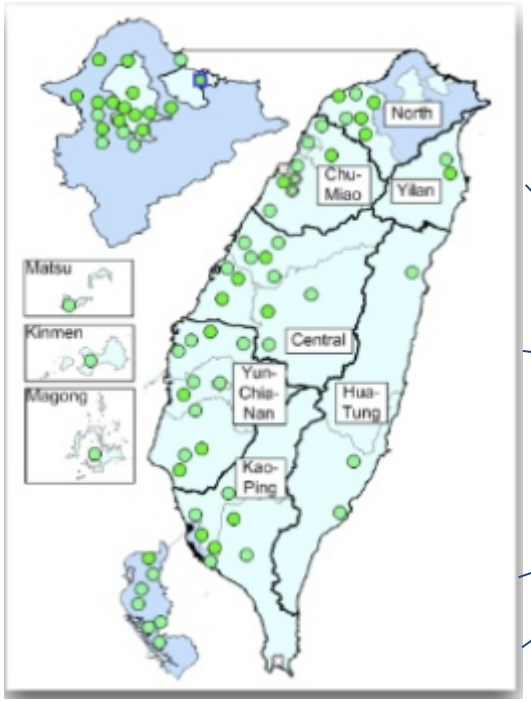
Indiana University Bloomington, USA

DATA TO INSIGHT CENTER

PERVASIVE TECHNOLOGY INSTITUTE

An approach to hands-on training for on-line courses drawing on

Day of data from generation to use



curation

SEAD

PID assignment

RDA RESEARCH DATA ALLIANCE

Handle.Net®

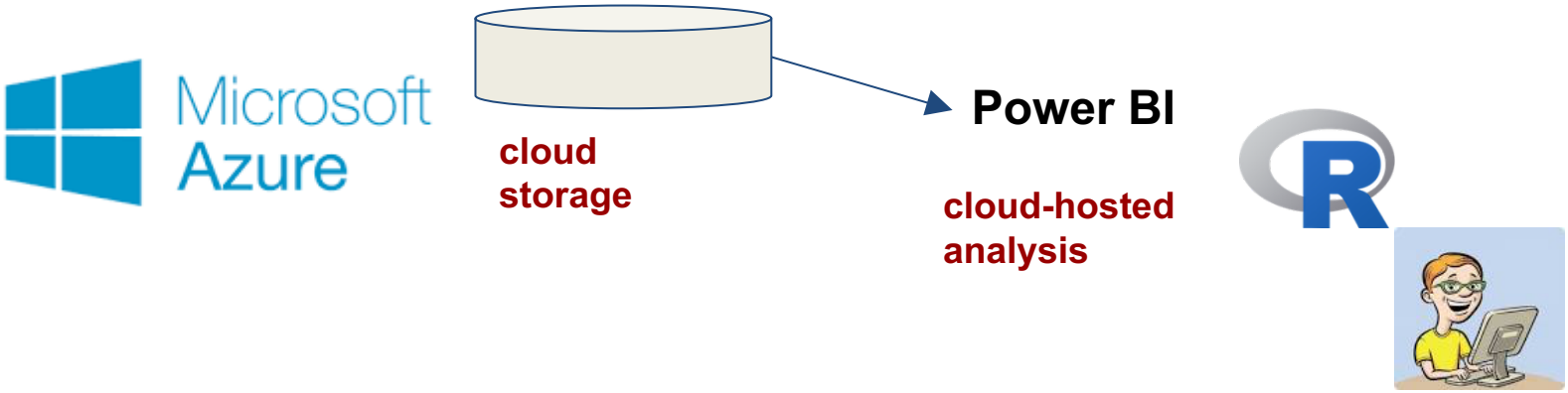
Microsoft Azure

cloud storage

Power BI

cloud-hosted analysis

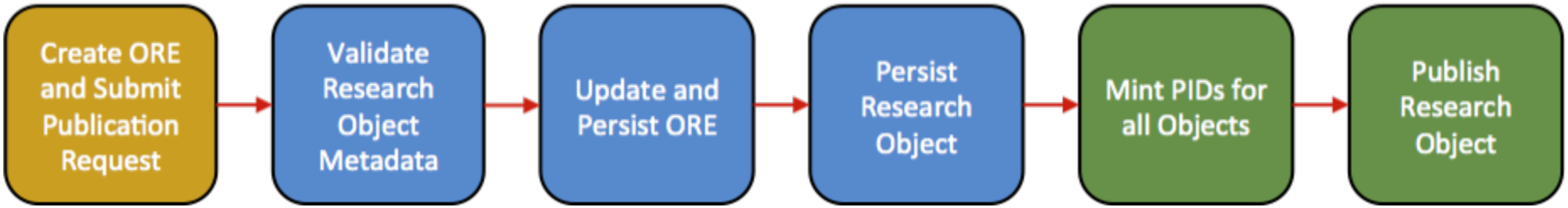
R



On-line courses in need of hands-on project component struggle to find easy to use resources, and struggle with not getting bogged down teaching platform access methods.

SEAD publication process

SEAD publishing model is one of publishing complex, heterogeneous Research Objects under a single DOI.

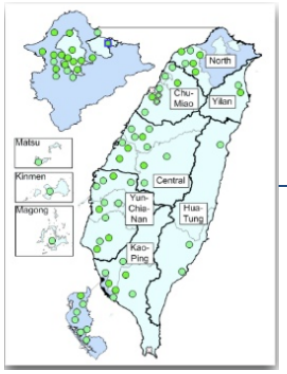


This hands on demo shows the conversion of RO == 1 DOI model to one where PIDs are pushed down into the RO

SEAD

Data model conversion

Data generated continuously by AirBox sensors.



curation

SEAD

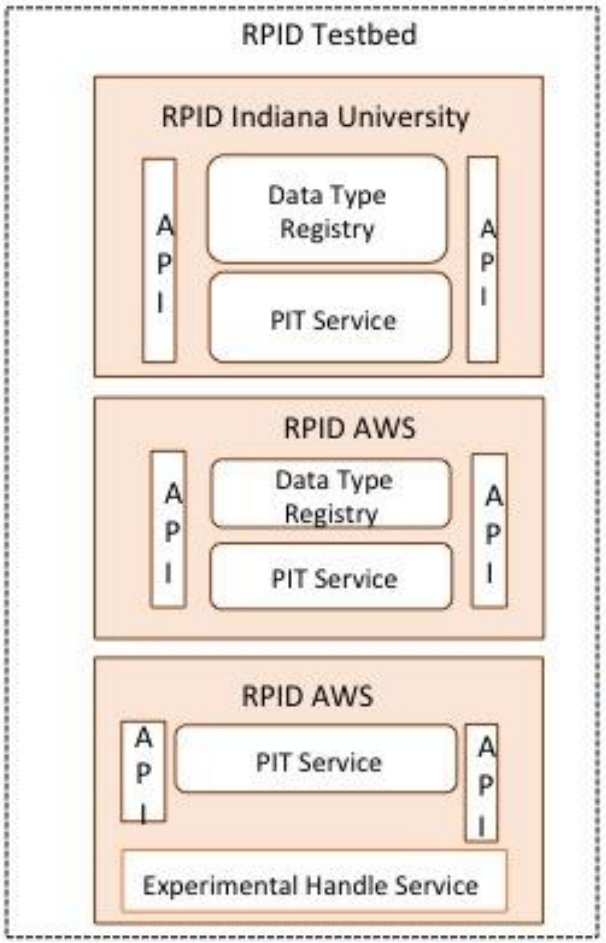
Data are lumped on per sensor/per day basis. Each sensor-day is considered a unique Research Object.

To conform to SEAD v2.0, a single OAI ORE map is defined for the entire RO

Sensor-day is then opened up and assigned PIDs down to the lowest granularity.

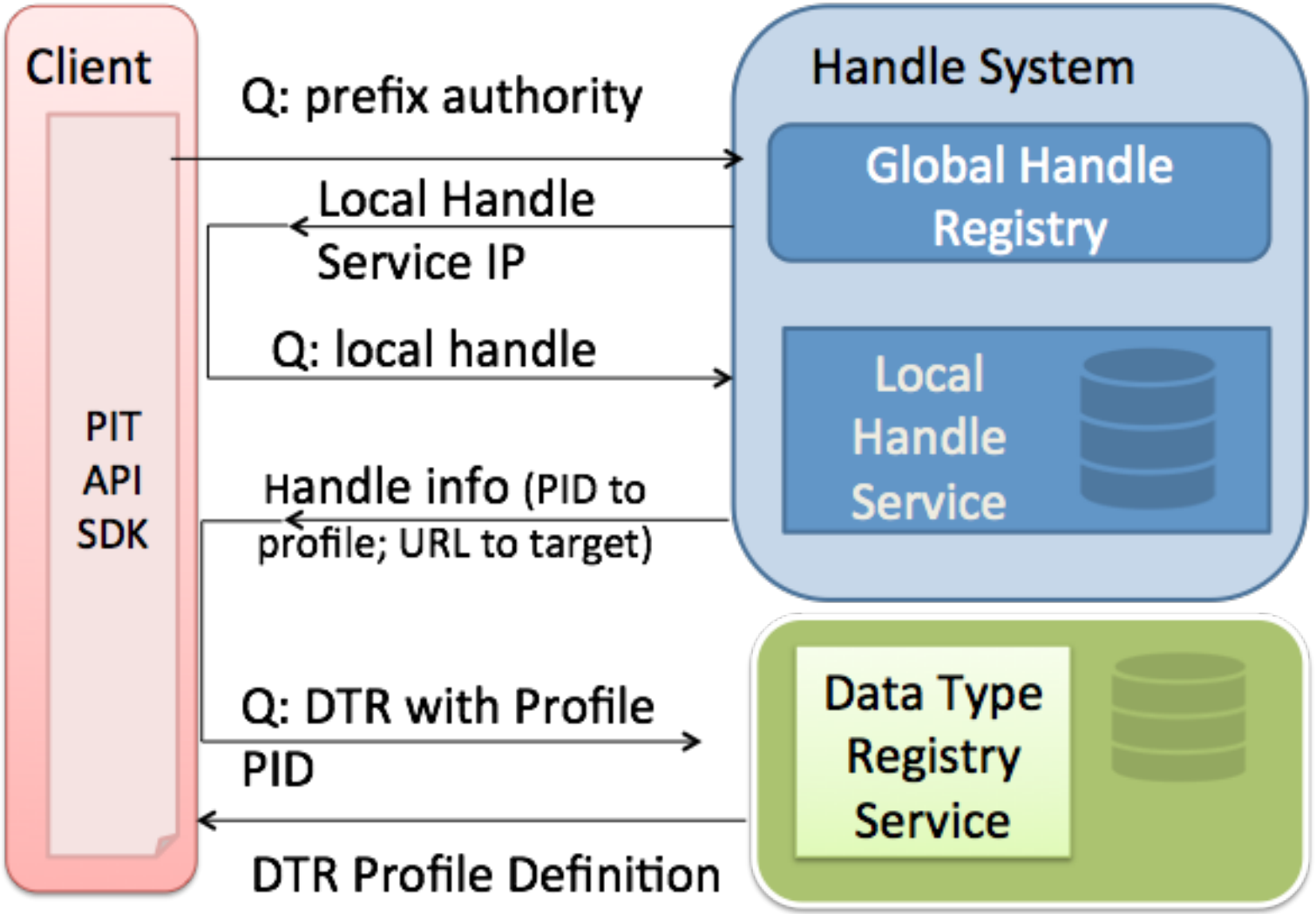
In this version of SEAD, RO level ORE map is used to maintain connections to atomistic objects

PID assignment using RPID ("rapid") Testbed



RPID testbed funded by the National Science Foundation under Award 1659301

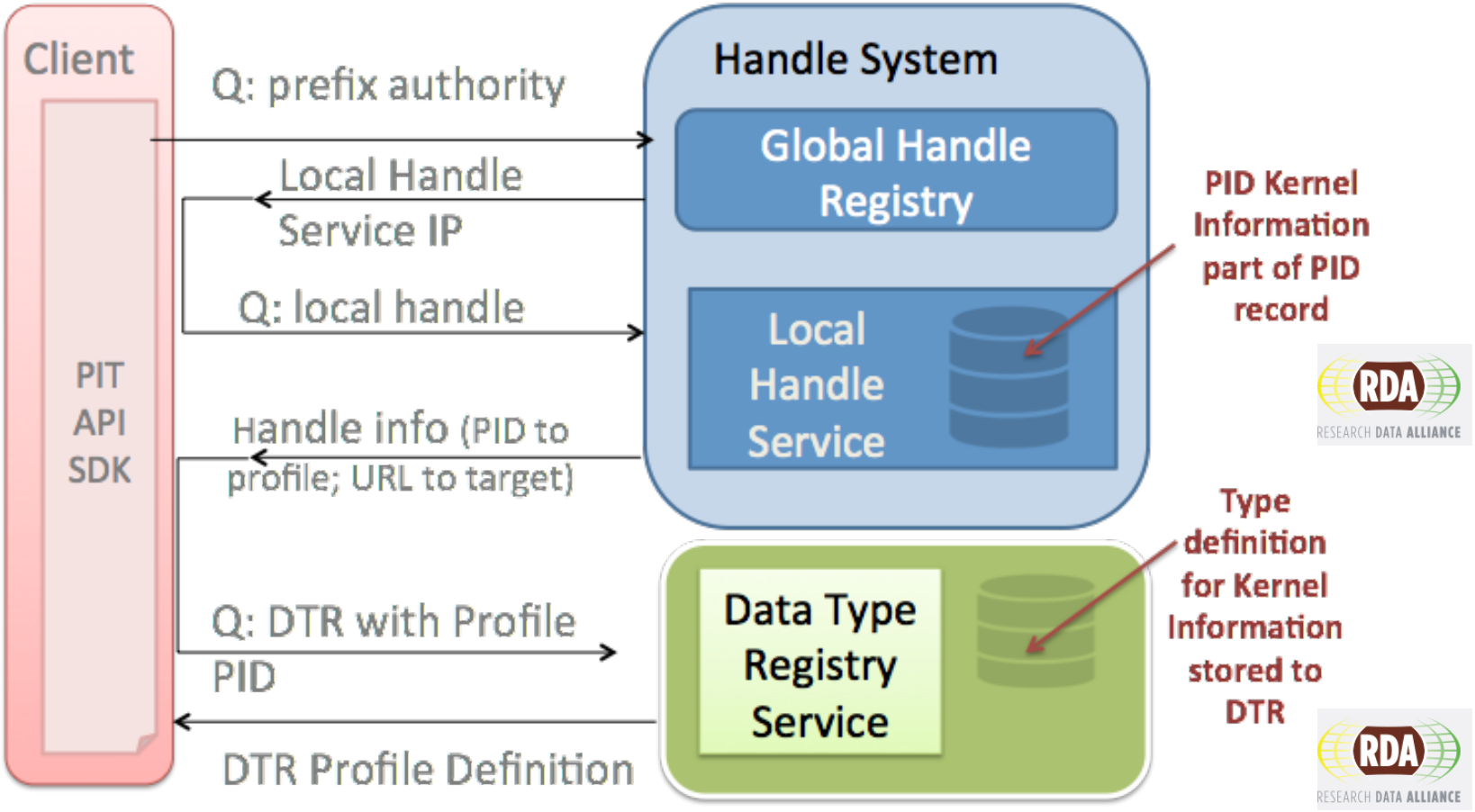
Handle System Resolution



PID Kernel Information

- The Handle system allows for inclusion of a small amount of information to be stored at the local handle server. We are exploiting this in an RDA PID Kernel Information WG to define what we call PID kernel information.
- PID Kernel Information, if defined wisely, can enable an entirely new ecosystem of data services operating at Internet speeds.

PID Kernel Information



Persistent IDs

- RPID testbed assigns only test temporary handles.
- All handles assigned by testbed are of type:
  - *11723.1.test, 11723.2.test, … 11723.8.test :* assigned for internal use
  - 11*723.9.test.<proj name>* : assigned to projects

Kernel Information

- defined by a profile stored to the Data Type Registry
- pointer to the type of the data object
- pointer to the data object

	Property identifier	Content format	Mandatory?	Explanation
1	PID	Handle	Yes	Global identifier for the object; external to the PID Kernel Information
2	RDAKIProfileType	Handle	Yes	Handle to the Kernel Information type profile; serves as pointer to profile in DTR. Address of *DTR federation expected to be global (common) knowledge.*
3	digitalObjectType	Handle	Yes	Handle points to type defn in DTR. The type of the object (this should always be the same for this type of data, but would distinguish it from other data types). Distinguishing metadata from data objects is a client decision within a particular usage context, which may to some extent rely on the digitalObjectType value provided.
4	digitalObjectLocation	URL	yes	Pointer to the content object location (pointer to DO)

Hands on Demo

- One month of Air Quality data is published into an Azure repository on per device/day basis
- Each device/day is assigned a PID as is each reading within the device/day Research Object
- From an Azure VM, a student invokes a script (a PID aware client) to query PIDs and download related data into their local VM.
- After analyzing data, the results could be re-published into Azure with assigned PIDs

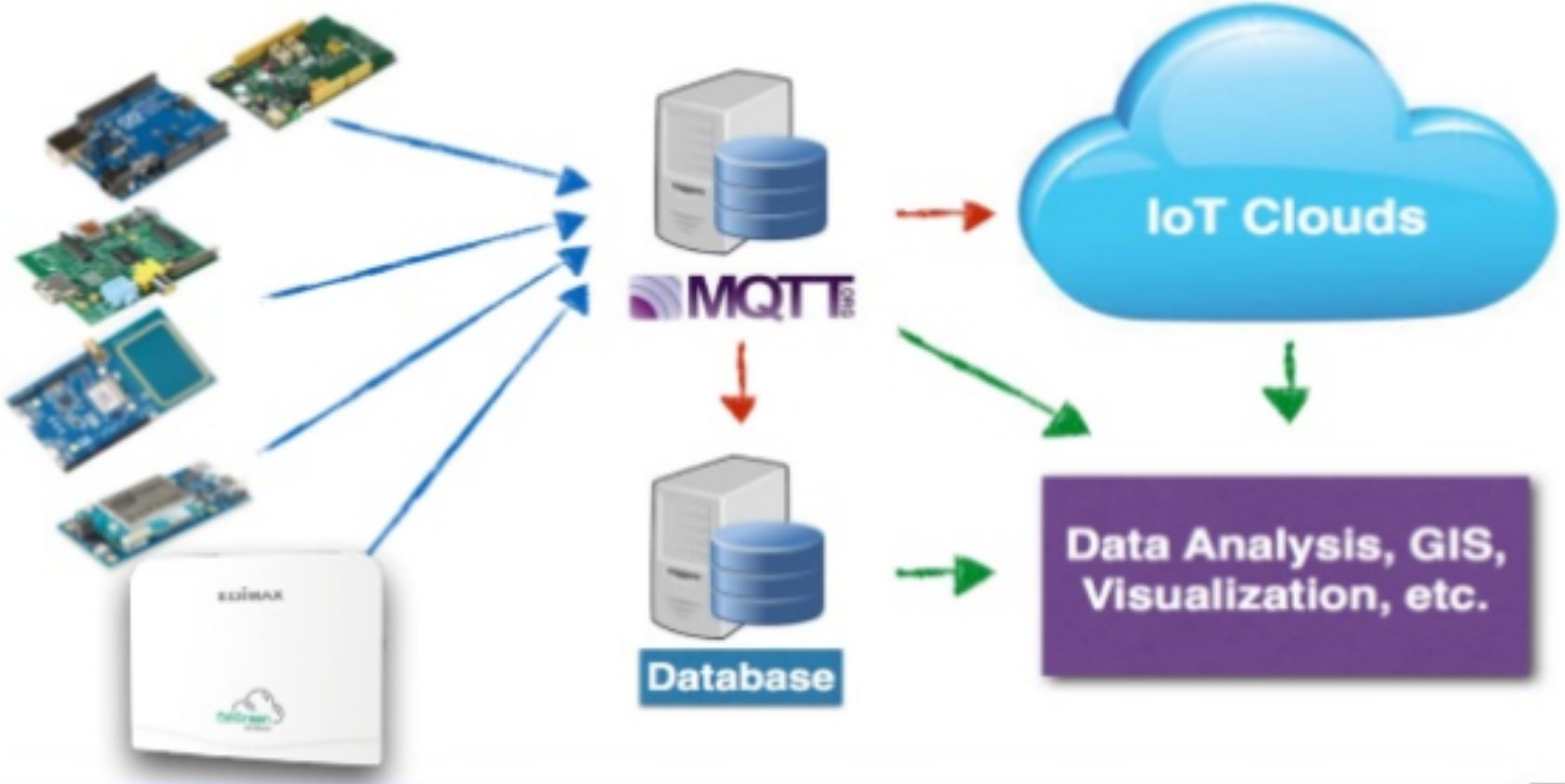
The Data

AirBox: a participatory ecosystem for PM2.5 monitoring

Ling-Jyh Chen, Academia Sinica
cclljj@iis.sinica.edu.tw

The ecosystem in a nutshell

The core spirit: open hardware/software/data



Raw Data

|ver_format=3|FAKE_GPS=1|app=MAPS|ver_app=5.1.5|
device_id=9E65F90B2642|date=2017-03-
30|time=06:31:55|device=LinkIt_Smart_7688_Duo|tick=704
098095|
s_t4=27.94|s_h4=50.51|s_b2=1009.24|s_d2=47|s_d0=68|s_
d1=84|d_t5=36.98|d_h5=33.51|gps_lat=23.445963|gps_lon
=120.490021|gps_fix=1|gps_num=15

Device id: 9E65F90B2642

Date: 2017-03-30 06:31:55

Sensor:
t : temperature sensor
d : dust sensor
h : humidity sensor
b: barometer

Device Data Fields

Some devices transmit data fields that others do not.

Shown are fields that most devices transmit.

Fields	Measurement & unit	Data type
device_id	12 character	string
date	year-month-day	date
time	hour:min:sec	time
device	name : LinkIt_Smart_7688_Duo	String:
s_t4	temperature : Celsius	float: %.2f
s_h4	relative humidity : %	float: %.2f
s_b2	barometer : [millibars]	float: %.6f
s_d2	dust sensor PM1: [ug/cm^3]	integer: 2 sig figs
s_d0	dust sensor PM2.5: [ug/cm^3]	integer: 2 sig figs
s_d1	dust sensor PM10 : [ug/cm^3]	integer: 2 sig figs
d_t5	device temperature	float: % .2f
d_h5	device humidity	float: %.2f
gps_lat	latitude	float : %.6f
gps_lon	longitude	float : %.6f
gps_fix	= 1	Integer
gps_num	# of satellites in gps fix = 15	Integer

Thanks to partners and sponsors

Accessing Azure Windows VMs

- Receive a username and password from us; ex:- seadtrain_demo1 and password
- Open the link, download to your laptop, open the file and click Connect . (**Mac users will need to download Microsoft Remote Desktop in App store**)
- Click Yes on the Security Certificate Confirmation screen