Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Technik & Architektur

ITC – Big Data Cloud

Schlussbericht

P. Zotter, E. Niederberger, R. Marek, Ch. Di-Battista, A. Lauber, R. Bärtsch

Horw, 13.10.2021

Inhaltsverzeichnis

1 Management Summary

Das ITC Projekt «Big Data Cloud» (ITC BDC) hat gezeigt, dass es an der HSLU dringend eine schlagkräftige und praxisorientierte Scientific IT braucht, welche diverse Dienstleistung anbieten kann, um die Forschenden und Studierenden besser zu unterstützen. In diesem Dokument sind, neben einer Zusammenfassung der Ergebnisse des ITC BDC, konkrete Empfehlungen, Anforderungen und Vorschläge für Angebote und Services einer Scientific IT zu finden. Eine solche Scientific IT ist aus Sicht der Autoren zwingend notwendig, damit die HSLU die Möglichkeiten der Digitalisierung, Internet of Things und Big Data optimal nutzen kann und die zukünftigen Anforderungen in Forschung und Entwicklung erfüllen kann.

2 Zusammenfassung

Im Zeitalter der Digitalisierung, Internet of Things, Industrie 4.0 und Big Data werden Forschende mit einer rasant wachsenden Datenmenge konfrontiert. Dies geht von der Datenerfassung, -übertragung, -speicherung, -verarbeitung und -visualisierung über die elektronische Publikation bis hin zum so genannten "Data Life Cycle Management". Diese Datenflut wie auch neuen Möglichkeiten der Datenanalyse transformieren die Arbeitswelt. Vielen Forschenden fehlen jedoch teilweise die notwendigen Kompetenzen oder finanzielle und zeitliche Ressourcen sich diese anzueignen oder als Service zu beziehen.

Daher haben sich Forschende verschiedener Institute und Departemente der HSLU zusammengeschlossen, um in einem ITC Projekt Bottom-up Lösungen für ihr gemeinsames Bedürfnis zu finden, Forschungsdaten auf cloudbasierten Plattformen zu speichern und diese effektiv, automatisiert und mit der Möglichkeit der Anwendung von künstlicher Intelligenz und Machine Learning, weiter zu bearbeiten, zu analysieren und zu visualisieren. Zudem wurde versucht einen groben Überblick über aktuelle Angebote und Möglichkeiten an der HSLU, sowie an anderen Universitäten/Hochschulen sowie kommerzieller Anbieter, zu geben.

Die Ergebnisse der Recherchen dieses Projektes haben gezeigt, dass es an der HSLU zwar teils Angebote, Services und Infrastruktur zu diesen Themenbereichen gibt, diese aber zu wenig bekannt, nicht direkt anwendbar und oft für das geringe Forschungsprojektbudget teuer sind. Zusätzlich ist auch die Kooperation und der Wissensaustausch innerhalb der HSLU zwischen verschiedenen Departementen, Instituten und Forschungsgruppen nicht optimal. Folglich laufen Entwicklungen parallel und/oder werden nicht weiterverwendet, was unnötige Ressourcen verschwendet. Zudem wird oftmals die Wartung von bestehenden Systemen unterschätzt und vernachlässigt. Desweitern sind aus Sicht der Autoren kommerzielle Big Data Plattformen nur bedingt geeignet für F&E Projekte an der HSLU und z.B. teilweise auch für SNF Projekte nicht zugelassen. Im Moment können Projekte mit sogenannten Bastellösungen noch durchgeführt werden, aber auch von Seite der Finanzierungs- und Industriepartner steigen die Anforderungen laufend. Zudem ist mit den aktuellen Lösungen der Kostenanteil für die Datenerfassung und Auswertung zu hoch. An anderen Hochschulen und Universitäten gibt es teils gute selbst entwickelte Lösungen bzw. umfangreiche Angebote, diese sind meist aber nur intern verfügbar und dementsprechend oftmals zu wenig detailliert dokumentiert, um diese nachbauen zu können.

Darum wurden in diesem Projekt zwei Musterlösungen für cloud-basierte Plattformen entwickelt, welche die Datenerfassung, -übertragung, -speicherung, -visualisierung und Weiterverarbeitung umfassen. Diese Musterlösungen wurden dokumentiert und werden allen Mitarbeitenden der HSLU zur Verfügung gestellt. Die Dokumentation enthält Codebeispiele, Installationsdateien und Anleitungen, die es auch Forschenden mit geringen IT Kenntnissen erlauben soll, in wenigen Schritten und geringem Zeitaufwand (<2h) die Beispiel-Cloud-Datenbanklösungen auf Ihren Systemen zu installieren.

Neben der Dokumentation in diesem Bericht sind Codebeispiele, Installationsdateien und Anleitungen unter folgenden URLs zu finden:

- Sensor-Data-Cloud (Datenerfassung, Speicherung/Bereitstellung für Analysen und Visualisierung von Zeitreihen)
  - https://gitlab.enterpriselab.ch/sensor-data-cloud/
  - https://www.hslu.ch/de-ch/hochschule-luzern/forschung/projekte/detail/?pid=4159
- Swisens (Lösung für Big Data Analysen)
  - https://swisens.ch/swisens-poleno/#data-explorer

In diesem Projekt wurden zwei Musterlösungen entwickelt, welche die Bedürfnisse der projektbeteiligten Forschenden nach einer einheitlichen und „State-of-the-Art" Lösung kurzfristig löst. Die längerfristige Wartung dieser Lösungen ist jedoch nicht sichergestellt. Die Musterlösungen können auch nicht alle verschiedenen Anwendungsgebiete abdecken, die an der HSLU benötigt werden. Aufgrund der genannten Punkte sowie den Erfahrungen und Recherchen zu bestehenden Services und Kooperation bzw. Wissensaustausch innerhalb der HSLU, empfehlen die Autoren dieser Studie der HSLU dringend, die Forschenden und Studierenden besser zu unterstützen und eine Scientific-IT aufzubauen, welche diverse Dienstleistungen anbieten kann. Ganz nach dem Motto: „Forschende sollen sich auf das konzentrieren was sie gut können - Forschen." Ohne einer Scientific IT wird es für viele Forschende an der HSLU schwierig die Möglichkeiten der Digitalisierung, Internet of Things und Big Data optimal zu nutzen und die zukünftigen Anforderungen in der Forschung und Entwicklung zu erfüllen.

Als Beispiel mit umfangreichen Services kann die Scientific IT der ETH Zürich genannt werden (https://sis.id.ethz.ch/). Es gibt an der HSLU bereits Projekte bzw. Bestrebungen in die Richtung einer Scientific IT. Zum Beispiel ist im Projekt Nr. 108682-00 „Forschungsdatenmanagement Umsetzung" der Aufbau eines Service Teams. Zudem gibt es konkrete Bestrebungen für eine stärkere Zusammenarbeit der Departemente Informatik und Technik & Architektur bei den Themen Internet of Things (IoT), Augmented Reality/Virtual Reality (AR/VR) sowie Robotik. In beiden Departementen gibt es für die verschiedenen Themen zuständige Personen, die über entsprechende Kompetenzen, Projekte und Angebote verfügen, welche kontinuierlich ausgebaut werden (siehe Inside Meldung „Zusammenarbeit der Departemente Informatik und Technik & Architektur" vom 13.07.2020).

Die Autoren dieses Berichtes unterstützen die Bestrebungen zum Aufbau eines IT Service Teams und die verstärkte Kooperation innerhalb der HSLU und empfehlen diese auf weitere Themen auszuweiten und alles mit der notwendigen Priorität, möglichst zeitnah mit genügend finanziellen Mitteln, umzusetzen. Nachfolgend werden Empfehlungen, Anforderungen und Vorschläge für Angebote und Services einer möglichen „Scientific-IT" aus Sicht der Autoren dieses Berichts beschrieben.

2.1 Empfehlungen, Anforderungen und Vorschläge für Angebote und Services einer Scientific IT

Aufgrund der Ergebnisse und Schlussfolgerungen dieser Studie, die im vorigen Kapitel zusammengefasst sind, empfehlen wir der HSLU dringend ein Scientific IT Team aufzubauen, welches den Forschenden bei verschiedenen Themen Support leisten, Infrastruktur bereitstellen und diese Systeme professionell unterhalten sowie Schulungen anbieten kann. Die Services der Scientific IT könnten durch die bereits bestehende HSLU IT, das Departement für Informatik sowie anderen Instituten und Forschungsgruppen mit den entsprechenden Kompetenzen, wie z.B. dem Institut für Elektrotechnik IET, geleistet werden. In einer zweiten, zeitnahen Phase soll eine Weiterbildungsoffensive für die Forschenden gestartet werden um möglichst viele, auch langjährige Mitarbeitende auf das minimal erforderliche Niveau zu heben.

Nachfolgend werden Anforderungen und Vorschläge für Angebote und Services einer schlagkräftigen und praxisorientierten Scientific IT an der HSLU aufgelistet.

- Allgemein (und für alle weiteren Punkte gültige Anforderungen und Services):
    - State-of-the-Art Knowhow und Support, von der Auswahl der benötigten IT-Lösungen über mögliches Software Development und Consulting bis hin zur Installation, dem Betreiben und dem Unterhalt
    - Best Practice Beispiele inklusive Programmcode und Installationsanweisungen
    - Standardisierte, kommerzielle und Open-Source Lösungen, die mit den Anforderungen von Finanzierungs- und Forschungspartnern konform sind
- Datenerfassung und Monitoring:
    - Empfehlung für IoT Devices
    - Datenerfassungshardware (Industriestandard-Messrechner aber auch neue günstige Möglichkeiten wie z.B. Raspberry Pi)
    - Tools und Software zur Datenerfassung sowie automatisierte Daten-Prüfung
    - Integrierte Überwachung und Alarmierung bei Fehlfunktion
- Datenübertragung:
    - Übertragungsmöglichkeiten (LoRaWAN, LTE, WLAN)
    - Kompetenzen aufbauen für verschiedene Übertragungsprotokolle und Gateways
    - Automatisiert mit Alarmierung bei Fehlfunktion
- Datenspeicherung inklusive Infrastruktur:
    - Consulting zu Datenmanagement allgemein
    - Zur Verfügung stellen von Infrastruktur
        - Virtuelle Maschinen und/oder physikalische Server
          (Hosting, Installation, Unterhalt, Wartung und Backup)
    - Web- bzw. Database Hosting inklusive ausreichendem Speicherplatz, automatischen Backups und Langzeit-Datenarchivierung
    - Alarmierung bei Fehlfunktion
    - Consulting zu verschiedenen Datenbanken (Einsatzmöglichkeiten, Funktionalitäten, Vor- und Nachteile)
        - Z.B. PostgreSQL, InfluxDB für Zeitreihendaten, Graph Datenbanken für Metadaten, etc.
- Datenbearbeitung, -auswertung, -visualisierung und -verwaltung
    - Bereitstellung und Unterstützung für Datenzugriff auf Datenbanken aus Python, R, Tableau, SPSS, Excel etc.
    - Webhosting von verschiedenen Angeboten, z.B.:
        - Dashboards (Grafana) inkl. Zugriff für Kunden und Forschungsteilnehmer
        - Python Flask Webpages
        - R Shiny für eigene Webpage basierte Applikationen oder Dashboards für R
    - Bereitstellung und Consulting bei der Erstellung von State-of-the-Art Tools und Programmcodes
        - Renku, Jupyter Notebook, Jupyter Lab
    - Codeanalyse
    - Consulting in Machine Learning
- Datensicherheit, Datenschutz und rechtliche Aspekte
    - Datensicherheit bei der Übertragung
    - Verschlüsselung
    - Fernzugriff
    - Datensicherheit bei der Speicherung

3 Projektbeschreibung

3.1 Ausgangslage

Mit der Digitalisierung der Forschung, von der Datenerfassung über die elektronische Publikation bis hin zum so genannten "Data Life Cycle Management" (DLCM), werden Forscher mit einer rasant wachsenden Datenmenge konfrontiert. Einerseits müssen die Forschenden den neuen Vorgaben des Forschungsdatenmanagements gerecht werden, die von Förderorganisationen und Verlagen gefordert werden, andererseits müssen die Forschenden der angewandten Wissenschaften das Know-how haben um Lösungen für die Wirtschaft anbieten zu können, die oft andere Anforderungen hat als die Wissenschaft (z.B.: Datensicherheit, Visualisierung, etc.). Die rasant wachsende Datenmenge und die explosionsartig wachsenden neuen Möglichkeiten der Datenanalyse transformieren die Arbeitswelt.

Oftmals, auch an der HSLU, werden aber noch veraltete Methoden angewendet, z.B. werden manuell CSV Dateien mit den Messdaten mittels Remote Zugriff vom Messgerät/-rechner kopiert und die Datenanalyse und Visualisierung wird manuell im Post-Processing und teilweise noch mit Excel, das für komplexere Aufgaben einfach limitiert ist, durchgeführt. Das ist zum einen darauf zurückzuführen, dass die Forschenden oftmals keine Zeit bzw. nicht ausreichend Budget haben, um neuere bessere Lösungen zu suchen und einzusetzen. Andererseits fehlt es aber teilweise auch am Willen, den Kompetenzen, oder schlicht dem Wissen über die neuen Möglichkeiten, aber es gibt auch Vorbehalte gegen diese neuen Ansätze. Zusätzlich zu diesen Punkten ist auch die Kooperation und der Wissensaustausch zwischen verschiedenen Departementen, Instituten und Forschungsgruppen innerhalb der HSLU oftmals nicht optimal und einige Entwicklungen laufen parallel, was unnötige Ressourcen verschwendet.

Am ITC Workshop zum Call im September 2018 zeigte sich, dass es ein zentrales gemeinsames Anliegen vieler Forschenden war, eine intelligente cloud-basierte Datenablage für Forschungsdaten zur Verfügung zu haben, mit der die Daten effektiv weiterverarbeitet werden können, auch mit der Möglichkeit künstliche Intelligenz und Maschine Learning anwenden zu können. Daher wurde im Rahmen des ITC „Digitale Transformation der Arbeitswelt“ das ITC Mantelprojekt «Big Data Cloud» (ITC BDC) mit drei Kooperationsprojekten durchgeführt.

3.2 Zielsetzung

Das Ziel des ITC BDC ist es den Umgang mit grossen Datenmengen, von der Datenerfassung über die Datenablage und Visualisierung bis hin zur Weiterverarbeitung und einen möglichen Umgang mit Metadaten, für Forschende der HSLU bedeutend zu vereinfachen. Im Zeitalter der Digitalisierung, Internet of Things und Industrie 4.0 sollten modernere und effizientere Lösungen angewendet werden, als manuell über Remotezugriff CSV Dateien von den Messrechnern zu speichern und dann, z.B. mit Excel, auszuwerten.

Darum werden in diesem Projekt mehrere Beispiele für cloud-basierte, intelligente Plattformen erstellt, die von der Datenerfassung, -übertragung, -speicherung, -visualisierung und Weiterverarbeitung umfassen, dokumentiert und allen HSLU Zugehörigen zur Verfügung gestellt. Die Dokumentation enthält Codebeispiele, Installationsdateien und Anleitungen inklusive Videos, die es auch Forschenden mit durchschnittlichen IT-Kenntnissen erlauben soll, in wenigen Schritten und geringem Zeitaufwand die Beispiel-Cloud-Datenbanklösungen auf Ihren Systemen zu installieren.

Ausserdem soll noch ein Überblick gegeben werden

- über weitere Projekte an der HSLU zum Thema Big Data
- zum aktuellen Angebot von Services und vorhandener Infrastruktur an der HSLU
- wie es andere machen
- Empfehlungen und Wünsche an eine möglichen Science IT an der HSLU

3.3 Involvierte Institute und Departemente

Zusammen mit dem ITC Mantelprojekt ITC BDC wurden drei weitere ITC Kooperationsprojekte eingereicht, die am Mantelprojekt mitarbeiten und ihre unterschiedlichen Kompetenzen und Bedürfnisse einbringen. Die im ITC BDC entwickelte cloud-basierten, intelligenten Datenablage wird dann in Teilprojekten eingesetzt. Die involvierten Institute, Kompetenzzentren (CC) und Forschungsgruppen sind in Tabelle 1 aufgelistet.

Tabelle 1: Departemente, Institute, Kompetenzzentren (CC), Forschungsgruppen und Personen, die an den verschiedenen ITC Projekten beteiligt waren.

Projektname	Projektbeteiligte	Departement HSLU	CC / Forschungsgruppe
ITC Mantelprojekt – Big Data Cloud	Erny Niederberger (administrativ weitergeführt von Armin Taghipour, CC Autonomous Systems & Robotics)	Institut für Elektrotechnik	CC Electronics
	Peter Zotter, Adrian Lauber	Institut für Maschinen- und Energietechnik	CC TEVT, Fachgruppe Bioenergie
	Rene Bärtsch		CC Mechanische Systeme
	Christian Di Battista	Institut für Elektrotechnik	CC Digital Energy & Electronic Power
	Reto Marek	Institut für Gebäudetechnik und Energie	Zentrum für Integrale Gebäudetechnik
	René Meier, Michael Handschuh, Tobias Merinat	Departement Informatik	Systems and Software Research Lab
ITC – Beflaggungstechnik 4.0	• Rene Bärtsch • Christian Jost	• Institut für Maschinen- und Energietechnik • Institut für Elektrotechnik	• CC Mechanische Systeme • CC Digital Energy & Electronic Power
ITC – Low Cost Emissionsüberwachung von Holzfeuerungen	• Peter Zotter & Adrian Lauber • Christian Di Battista	• Institut für Maschinen- und Energietechnik • Institut für Elektrotechnik	• CC TEVT, Fachgruppe Bioenergie • CC Digital Energy & Electronic Power
ITC – Digitale Schädlingsbekämpfung in der Landwirtschaft	• Erny Niederberger • Prof. Dr. Marc Pouly	• T&A Institut für Elektrotechnik • Departement Informatik	• CC Electronics • Algorithmic Business Research Lab

3.4 Weitere Projekte und Massnahmen an der HSLU

3.4.1 Projekt Forschungsdatenmanagement – Umsetzung

Seit Anfang 2020 läuft an der HSLU das Projekt Forschungsdatenmanagement – Umsetzung (Projektnummer:108682-00, https://inside.hslu.ch/fs/ba/InfoDocsBA/Projektauftrag%20_Forschungsdatenmanagement%20Umsetzung.pdf), das sich ebenfalls mit Anforderungen an das Datenmanagement und unter anderem auch technischen Themen (Infrastruktur, Tools, Schnittstellen, usw.) und bestehenden nationalen IT Infrastrukturen beschäftigt. Michael Wanner, der Projektleiter dieses Projekts, hat das ITC

BDC auch kontaktiert und die Projektbeteiligten des ITC BDC haben einen Fragenkatalog zur Datenerhebung beantwortet. Die Fragen und Antworten sind im Anhang (Kapitel 7.1) zu finden.

Bisher sind aus dem Projekt Forschungsdatenmanagement – Umsetzung noch keine konkreten Resultate verfügbar, als Fazit kann jedoch genannt werden, dass das Thema Forschungsdatenmanagement bei der HSLU Leitung präsent ist. Gemäss Michael Wanner hat die Projektsteuerung am 10.12.2020 folgendes entschieden:

- Aufzeigen von Varianten zum Aufbau einer IT-Plattform mit GIT, GIT-Lab und Renku bis Ende März 2021
- Aufzeigen eines Angebotes von Plattformen, um Forschungsdaten und -dokumente zu speichern, teilen und archivieren, die Renku nicht benötigen bis Ende März 2021
- Aufzeigen der Dienstleistungen, welche von einem Service-Team bei IT Services angeboten
- werden können bis Ende März 2021
- Erstellung einer Informationsplattform für Forschende (1. Bei Grants Office, 2. Thema
- Forschungsdaten durch Bibliotheken)
- Richtlinie bis Ende April 2021 finalisieren
- Textbausteine können erst erfolgen, wenn die technischen Plattformen klar sind

Aufgrund der ersten 3 Punkte, hat das ITC BDC entschieden Empfehlungen und Wünsche an eine möglichen Science IT an der HSLU an die Projektleitung des Projekts «Forschungsdatenmanagement – Umsetzung» bis Ende März 2021 zu übermitteln, die in Kapitel 1 zu finden sind.

3.4.2 Verstärkte Zusammenarbeit der Departemente Informatik und Technik & Architektur

In einer Inside Mitteilung vom 13. Juli 2020 wurde eine verstärkte Zusammenarbeit der Departemente Informatik und Technik & Architektur bei den Themen Internet of Things (IoT), Augmented Reality/Virtual Reality (AR/VR) sowie Robotik angekündigt. In beiden Departementen gibt es für die verschiedenen Themen zuständige Personen, die über entsprechende Kompetenzen, Projekte und Angebote verfügen, welche kontinuierlich ausgebaut werden (Personal, Infrastruktur). Beide Departemente nutzen dieses Know-how in Ausbildung, Forschung sowie Weiterbildung und teilweise auch für Dienstleistungen. Der Austausch in diesen Bereichen kann für beide Departemente einen Mehrwert darstellen, da damit die Ressourcen zusätzlich gebündelt werden können. Jedes Departement stellt für die Bearbeitung dieser Themen ein Stundenbudget von initial 100h zur Verfügung.

Diese verstärkte Zusammenarbeit ist sehr zu begrüssen, da die Erfahrungen in diesem Projekt mit der Informatik gezeigt haben, dass die Ressourcen für die Erfüllung interner Dienstleistungsaufträge seitens der Informatik zu gering sind (zeitlich und finanziell) und daher Möglichkeiten und auch der Wille zur verstärkten Zusammenarbeit der Informatik mit anderen Abteilungen der HSLU eingeschränkt sind. Eine frühzeitige Planung und Miteinbeziehung des Departements für Informatik zusammen mit einem ausreichenden Budget ist sicher sehr hilfreich. Es wäre aber zu begrüssen, wenn die Zusammenarbeit auf weitere Themen und Institute, z.B. IGE, IME und IET, ausgedehnt werden würde.

4 Aktuelle Services und Infrastruktur an der HSLU

4.1 Services

4.1.1 HSLU IT Services

An der HSLU gibt es einen eigenen IT Service, welcher dafür sorgt, dass Computer, Netzwerke, Applikationen, Datenbereitstellungen, Email und mehr funktionieren. Mit den Abteilungen Engineering und Application Development kann der HSLU IT Service auch bezüglich dem Thema BigData

Dienstleistungen erbringen. Dies kann von Projekt-Supporthilfen bis zu managed Servern reichen. Konkrete Dienstleistungen im Sinne einer Science IT werden jedoch nicht aufgelistet und es müssten jeweils projektspezifische Anfragen gestellt werden und deren Bearbeitung zur Gänze aus den Projektmitteln der Anfragenden beglichen werden müssen.

Das HSLU IT Service ist auch zuständig für das Freischalten von benötigten Ressourcen bei SWITCH (siehe Kapitel 4.1.3) und deren Verrechnung.

4.1.2 zhb - Zentral- und Hochschulbibliothek Luzern

Auf der Homepage der zhb gibt es zahlreiche Informationen zum Umgang mit Forschungsdatenmanagement. Zudem wird eine kostenlose Beratung für effizientes Forschungsdatenmanagement und Unterstützung bei der Erstellung eines Datenmanagementplans bzw. dessen Korrektur angeboten.

Am 27.5.2020 fand ein Webinar der zhb zum Forschungsdatenmanagement statt, an dem der Projektleiter des ITC BDC teilgenommen hat. Folgendes Fazit wurde gezogen:

- Das Forschungsdatenmanagement, so wie es am Webinar vorgestellt wurde, ist nicht praxisbezogen genug für Projekte der Angewandten Forschung und Entwicklung.
- Es gibt keine konkret verfügbare Lösung für die Ablage von Big Data

4.1.3 Switch

SWITCH ist eine Schweizer Stiftung, die seit 1987 Betreiberin des Schweizer Wissenschaftsnetzes der Hochschulen ist und bietet dynamische Rechen- und Speicherdienste an. Diese Dienste können von Hochschulen (auch alle Angehörige der HSLU) aber auch von Firmen und Einzelpersonen in der Schweiz genutzt werden. Für Cloud Anwendungen bietet vor allem SWITCHengines interessante Möglichkeiten, z.B. dynamische Rechen- und Speicherdienste in Form von virtuellen Maschinen, die beliebig (Anzahl CPUs, RAM, Speicher, Adressen) konfiguriert werden können.

Um SWITCHengines nutzen zu können, muss man sich zuerst dort registrieren. Die benötigten Ressourcen für die virtuelle(n) Machine(n) müssen dann über den HSLU IT-Helpdesk beantragt werden. Im Moment werden von der HSLU IT die Kosten noch nicht an die Forschenden weiterverrechnet. Das kann sich in Zukunft aber noch ändern. Informationen zu den aktuellen Preisen sind hier zu finden.

Wie man eine virtuelle Maschine auf SWITCHengines einrichtet, wird in der Dokumentation zur «Sensor Data Cloud» (Kapitel 6.2) unter https://gitlab.enterpriselab.ch/sensor-data-cloud/mdc/-/blob/master/documentation/SWITCHengines/README.md beschrieben.

4.1.4 Enterprise Lab

Das Enterprise Lab des Departments für Informatik ist als Rechencenter mittlerer Grösse aufgebaut und implementiert die neusten Technologien wie Datacenter Network, Storage Area Network, Block- und File-Speicher sowie x86 Server und Mainframe. Ursprünglich wurde es entwickelt, um den Studierenden modernste Computerressourcen zur Verfügung zu stellen, welche sie für alle möglichen Anwendungen und Übungen brauchen. Mittlerweile wurde es aber zu einem cloud-basierten Dienst für Forschende weiterentwickelt als «Infrastructure as a Service (IAAS)». Es können jegliche Netzwerke, Server und Speichersysteme konfiguriert werden mit der Sicherheit der virtuellen Firewall, die von Checkpoint bereitgestellt wird. Von VMware werden Tools für die Cloud angewendet und Cisco stellt die Nexus 1000 Technologie zur Verfügung, mit der virtuelle Switches konfiguriert werden können. Es ist jedoch nicht klar, ob die angebotenen Services auch über längere Zeiträume genutzt werden können, da die bereitgestellten Ressourcen bis anhin meistens nur Projektspezifisch eingesetzt wurden.

4.2 Infrastruktur

Gemäss den Erfahrungen der Projektbeteiligten erarbeiten am Departement Technik und Architektur verschiedene Institute und die Kompetenzzentren jeweils eigene individuelle Lösungen.

Beispiele dazu sind:

- iHomeLab: eigene Server-Infrastruktur und Plattformen
- IET: eigene Infrastruktur auf Enterprise-Lab
- IGE: eigene Infrastruktur testweise auf kleinem NAS (Bild 1), sonst Daten auf Netzlaufwerk im Projektordner. Bei Geheimhaltung Zugriffseinschränkung auf dieser Ebene.
- IME:
    - Infrastruktur vom IET
    - einige CCs auch ähnlich wie IGE mit den Daten auf dem HLSU Netzlaufwerk. Die Daten werden zudem meist manuell vor Ort oder manuell per Remote Zugriff vom Messgerät/Messrechner aufs Netzlaufwerk kopiert und die Datenauswertung im Postprocessing durchgeführt.
    - Das CC Fluidmechanik und Hydromaschinen verfügt über große Erfahrung auf dem Gebiet der CFD-Simulationen und Big Data. Das Institut verfügt sowohl über interne als auch kommerzielle CFD-Löser mit entsprechenden Pre- und Postprocessing-Utilities sowie über zwei HPC-Cluster (900 Cores, 5400 GB RAM, OmniPath, 40 TB SSD und 900 Cores, 1800 GB RAM, Infiniband, 112 TB Speicher), die die Simulation großer und sehr komplexer Fälle ermöglichen.
- HSLU Informatik ABIZ: ABIZ investierte kürzlich in sechs selbstverwaltete Workstations mit jeweils zwei hochmodernen GPUs (z.B. NVIDIA Tesla P100), die sich gut für das Training von tief lernenden Modellen für die Bildverarbeitung eignen. Diese Maschinen sind im Vergleich zu gleichwertigen Cloud-Ressourcen auf Abruf deutlich kostengünstiger.

Wie bereits oben (siehe Kapitel 4.1.4) gibt es am Department für Informatik für das Enterpriselab ein Rechenzentrum, über das, so wie bei SWITCH auch (siehe Kapitel 4.1.3), Ressourcen für VMs und andere Cloud Dienste bezogen werden können. Diese sind jedoch vielen Forschenden nicht bekannt oder es fehlen die Kompetenzen für die Anwendung, da es wenig bis keinen Support gibt.

Ausserdem können bei der HSLU IT gemanagte Server beantragt werden, die jedoch relativ teuer sind. Für einen Medienserver zum Betrieb eines Wikis musste das ITC BDC für 1 Jahr rund 5000 CHF an die HSLU IT bezahlen.

Bild 1: NAS mit Zeitreihen-Datenbank im Messmittelraum des IGE

Fazit

- Es gibt viel parallele und redundante Lösungen und es findet zwischen den Instituten und CCs wenig bis kein Erfahrungsaustausch statt, was sehr ineffizient ist.
- Viele Forschende beschäftigen sich mit Sicherheitsfragen und IT-technischen Herausforderungen, was unnötige Ressourcen verschlingt.
- Die von SWITCH und Enterpriselab angebotenen Services sind vielen Forschenden zu wenig bekannt und oft fehlt es auch an den Kompetenzen diese einzusetzen.
- Die möglichen Angebote der HSLU IT zum Thema Big Data sind ebenfalls nicht ausreichend bekannt, sind aber auch relativ teuer.
- Die eben genannten Punkte führen dann oftmals zu nachfolgendem negativem Beispiel (Bild 1).

5 Ist-Zustand, Angebote, Services, Infrastruktur anderer Hochschulen und Universitäten

5.1 EnhanceR

EnhanceR ist ein national und international anerkanntes Netzwerk für Schweizer Forschungs-IT-Kompetenz. Ziel des Vereins ist es, die Forschungsexzellenz in der Schweiz zu fördern, um die Führungsposition der Schweiz zu sichern. Er erreicht dieses Ziel durch den Zusammenschluss von Forschungs-IT-Fachgruppen an verschiedenen akademischen Institutionen in der Schweiz. Er schafft Mehrwert, indem er sein Fachwissen dort einsetzt, wo es am dringendsten benötigt wird, im Interesse der Gemeinschaft der Anwender und der Support-Teams von Scientific Computing-Anwendungen auf nationaler und internationaler Ebene. Er ist die juristische Person, die die Ergebnisse des 2019 abgeschlossenen EnhanceR-Projekts von swissuniversities weiterführt.

Ab 2020 zählt er 11 Schweizer Hochschul-, Infrastruktur- und Forschungsinstitutionen als Mitglieder. Die Mitgliedschaft ist gemäss den Statuten offen. Mit dabei von der HSLU sind Prof. Dr. René Hüsler und Dipl. Ing. Bruno Joho.

Die Projektgruppe hatte am 9.4.2020 eine Videokonferenz mit Thomas Wüest, Plamada Andrei Valentin, Chadha Tarun von EnhanceR, die Mitarbeiter der Scientific IT Services der ETHZ sind (Details zu den Personen sind hier zu finden: https://ethz.ch/en/the-eth-zurich/organisation/departments/ informatikdienste/personen/scientific-it-services-a-z.html).

Fazit:

Diese Spezialisten sind für 1000 CHF pro Tag verfügbar. Für einen konkreten Bedarf mit grösserem Aufwand können diese auch über einen Antrag und Vertrag gebucht werden. Dieses Vorgehen entspricht nicht einem besonders einfachen Zugang, allerdings für eine komplexere Fragestellung ist es sicher empfehlenswert EnhanceR zu kontaktieren.

5.2 HESO

Kontaktadresse

Univ. of applied Sciences, West. Switzerland, hepia
https://www.hes-so.ch

Dr. Nabil Abdennadher
Full Professor
Head of inIT research institute
Head of LSDS research group

Über Kontakte via EnhanceR kam ein Meeting zwischen der HESO und der HSLU zustande, worin Projekte gegenseitig vorgestellt wurden und mögliche künftige Zusammenarbeiten besprochen wurden.

Fazit:

Eine Zusammenarbeit mit der HESO wäre im Bereich von Forschungsprojekten sehr spannend, da sie ähnliche Projekte im Gebäude verfolgen wie das Institut für Gebäudetechnik und Energie der HSLU und seitens IT ähnliche Herausforderungen haben. Im Rahmen des vorliegenden Projektes wurde der Kontakt jedoch nicht weiter aufrechterhalten, da sie über eigene Infrastruktur verfügen von der die HESO keine weiteren Infos preisgeben resp. austauschen wollte. Am Schluss lief es auf eine Bezahlung nach Aufwand mit einem 50% Rabatt heraus, wenn wir mit Ihnen Informationen austauschen wollten. Dies entsprach nicht der Zusammenarbeit welche wir suchen.

5.3 Kollaboration ETHZ/EAWAG

Kontaktadresse

ITS Scientific SW & Data Management
ID Scientific IT Services
Dr. Uwe Schmitt

Über Kontakte via EnhanceR kam ein Meeting zwischen einem Software-Entwickler der ETH Scientifix IT Services zustande. Er entwickelte und unterhält für die EAWAG eine Python-Basierte Zeitreihen-Datenbank Infrastruktur, wo Forschende Ihre Feld-Messergebnisse hochladen und mit Metadaten versehen können. Dokumentation datapool: https://datapool.readthedocs.io/en/latest/

Fazit:

Die EAWAG und ETH plant längerfristig, die Infrastruktur neu zu implementieren und wäre an einer gemeinsamen Entwicklung interessiert. Der Kontakt hier sollte sicherlich aufrechterhalten werden.

5.4 Beispiele und Services anderer Hochschulen

- Das Physikdepartement der ETHZ hat beispielsweise ein IT Services Group (https://sis.id.ethz.ch/) mit breitem Angebot von Informationen bis hin zu konkreten Codebeispielen, Plattformen, Hardware Ressourcen und Services, die man beziehen kann. Nachfolgend ist der Service Katalog aufgelistet:
  - High Performance Computing
  - Software Development
  - Scientific Computing & Data Co-Analysis
  - Scientific Visualization
  - Data Science & Machine Learning
  - Research Data Management
  - Confidential Research Data
  - Consulting & Training
- Uni Freiburg hat «IT im Dienst von Forschung und Wissenschaft»
- An der Berner Fachhochschule gibt es eine IoT Architektur, die für Sensoren mit Zeitreihendaten einsetzbar ist und für verschiedene Anwendungen eingesetzt wurde (https://iot.i3s.bfh.ch/en/). Diese IoT Architektur ist praktisch ident mir der Time Series Database Lösung, die in diesem Projekt angewendet wurde (siehe Kapitel 6.2).

5.5 Kommerzielle Big Data Cloud Plattformen

Um eine Übersicht über Angebote und Möglichkeiten zu Big Data Plattformen von kommerziellen Cloud Anbietern zu erhalten wurde eine Recherche von Tobias Merinat durchgeführt und die Resultate am 19. August 2019 den Mitgliedern des ITC BDC präsentiert. Die Präsentation ist im Anhang (siehe Kapitel 7.2 im Anhang) zu finden.

Fazit aus der Recherche und unseren Erfahrungen

- Die Übersicht beinhaltet die Ansätze der grossen kommerziellen Anbieter. Für die konkreten Probleme in der Praxis bei der Forschung und Entwicklung an der Hochschule scheinen diese Serviceprovider nicht geeignet da kein Pilotprojet diesen Weg gegangen ist.
- Aus Rückmeldungen von Leuten des Enhancer Teams besteht bei der ein Nachteil, dass man die Daten praktisch nicht mehr von den Servern der Anbieter entfernen bzw. migrieren kann. Das heisst, wenn man diese für längere Zeit verfügbar halten will, muss man zahlen. Das wird dann für eine Langzeitarchivierung sehr teuer. Die ETHZ betreibt eigene Rechnercluster. Das hat immer noch Vorteile gegenüber kommerziellen Anbietern.
- Eine weitere Problematik ist der Ort der Datenspeicherung. Hier geben sich viele offene Fragen betreffend dem Datenschutz, der Datensicherheit und Anforderungen von Industriepartnern aber auch des Gesetzgebers.
- Die grossen Anbieter scheinen sich zur Zeit nicht für F&E Projekt an der HSLU zu eignen.

6 Ergebnisse – Musterlösungen für cloud-basierte Datenablagen

In diesem Kapitel werden die zwei Musterlösungen beschrieben, die in diesem Projekt erstellt wurden. Die «Swisens Lösung» wurde von der Firma Swisens AG in Kooperation mit der HSLU entwickelt. Neben der Beschreibung dieser Lösung finden sich auch Verweise auf weiterführende Informationen im nachfolgenden Kapitel. Die Musterlösung «sensor-data-cloud» wurde ausschliesslich an der HSLU entwickelt und die gesamten Informationen, Programmcodes und auch Installationsfiles werden allen Angehörigen der HSLU frei zugänglich zur Verfügung gestellt.

6.1 Swisens Ecosystems

Das Beispiel der Swisens AG zeigt an, wohin es in Zukunft gehen wird. Ein Messinstrument alleine genügt nicht um konkurrenzfähig zu bleiben. Es braucht Komplettlösungen für das BigData Management und eine Abstraktion für die Entwicklung von AI basierten Algorithmen, so dass die Kunden selber ohne Spezialwissen in der Lage sind die grossen Datenmengen zu bewältigen und das Potenzial von AI zu nutzen.

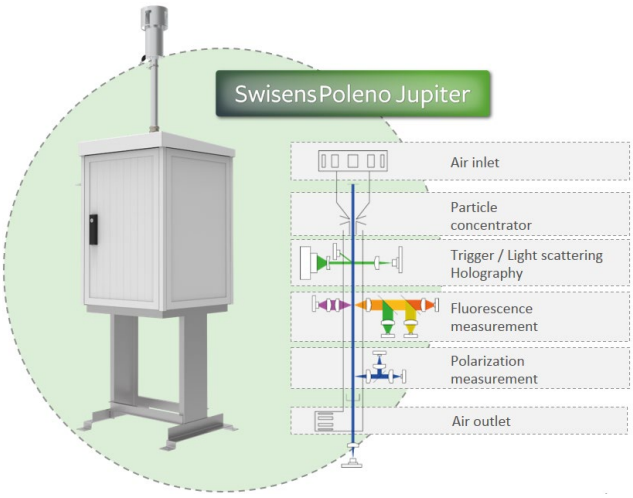


Bild 2: Swisens Echzeit Aerosopartikelmesssystem

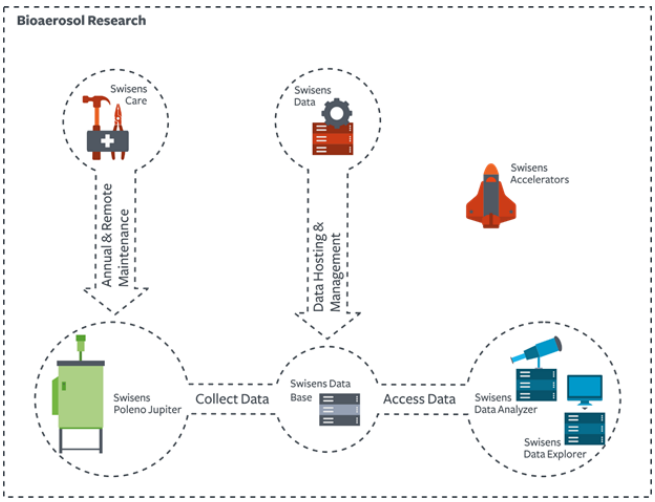


Bild 3: Übersicht der Komponenten des Swisens Ecosystem Bioaerosol Research

Die Swisens Komplettlösungen zeigen den Weg in die digitale Partikelmessung und Partikelidentifikation für Forschung und Industrie – mit Swisens Poleno und der führenden Echtzeit-Klassifizierung von Partikeln. Swisens Software ermöglicht die Verbindung von Benutzer, System und Daten. Mit den Swisens Softwarekomponenten kann man Daten analysieren, Messergebnisse auswerten, das Messnetz überwachen und maschinelles Lernen anwenden. In die Swisens-Softwarekomponenten fliesst die gesamte Erfahrung im Umgang mit BigData, künstlicher Intelligenz und dem Betrieb von Echtzeit-Messsystemen ein. Dabei bleibt Swisens der Philosophie eines offenen Datenzugangs, der Verwendung von Open-Source-Technologien und maximaler Freiheit für die Kunden treu.

Ein SwisensPoleno Jupiter Messinstrument liefert pro Tag Gigabytes von Messdaten. Darum hat Swisens eine Datenserver Software entwickelt welche die Daten von mehreren Messinstrumenten synchronisiert. Die Datenvisualisierung, Datenanalysen, Datenmanagement, die Überwachung der Messinstrumente sowie die Erweiterung der Erkennungsalgorithmen wird durch den SwisensDataExplorer ermöglicht. Für Wissenschaftler die eigene Visualisierungen und Analysetools nutzen wollen, bietet Swisens den SwisensDataExplorer an, damit kann einfach die Funktionalität des SwisensDataExplorere erweitert werden. Der SwisensDataAnalyzer ist unter einer OpenSource Lizenz.

Informationen dazu sind auf der Website von Swisens zu finden (swisens.ch) und auf deren YouTube Kanal (https://www.youtube.com/channel/UChrfjY4eeI7T2xeyquLStjA).

6.1.1 Swisens Komponenten

In diesem Kapitel wird eine kurze Übersicht über die relevanten Komponenten, die Swisens entwickelt hat, und wichtige Überlegungen zum Design des Systems angeführt.

SwisensDataExplorer

Der SwisensDataExplorer ist eine browserbasierte Software mit hilfreichen Tools zur Überprüfung von Messergebnissen und zur Überwachung von Hardwarekomponenten im SwisensEcosystem. Ob für ein einzelnes Messsystem oder ein Netzwerk, SwisensDataExplorer sorgt für einfache und schnelle Analysen der Messdaten und Betriebsparameter und ebnet den Weg für den eigenständigen Umgang mit maschinellem Lernen für alle Anwenderklassen.

Weitere Merkmale

- Untersuchung von Zeitreihen von Messdaten und Identifikationsergebnissen
- Histogramm-Auswertung auf Basis der Partikelgröße
- Überwachung von Betriebsparametern und Abfrage des Systemstatus
- Analysieren von aktuellen und historischen Zeitreihen, Partikelkonzentrationen und Datensätzen

Vorteile

- Sofortiger Zugriff
  - Die gemessenen Partikel und Messdaten können in Sekundenschnelle abgerufen, visualisiert und browserbasiert abgerufen werden.
- Einfache Datenanalyse
  - Die integrierten Analysewerkzeuge ermöglichen eine schnelle und effiziente Datenauswertung. Sowohl lokal als auch über Fernzugriff.
- AI-Schnittstelle
  - Datensätze können verarbeitet werden. Machine-Learning-Modelle können trainiert werden. Dies alles geschieht auf einer einheitlichen Platform.
- Quelloffen
  - Lizenziert nach dem GLP-Standard, wodurch eigenständige Optimierungen und erweiterte Anwendungsbereiche ermöglichen werden.

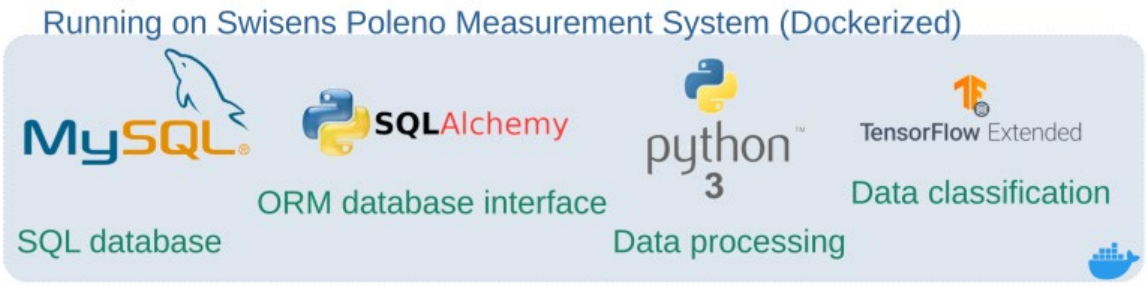


Bild 4: Verwendete Software Librarys im SwisensDataExplorer

SwisensDataAnalyzer

SwisensDataAnalyzer ist ein Toolset, welches eine effiziente Tiefenanalyse großer Datenmengen aus dem Messsystemen ermöglicht. Für fortgeschrittene Datenanalyse oder aussagekräftige Datenvisualisierung bietet SwisensDataAnalyzer eine übersichtliche und plattformunabhängige Arbeitsumgebung auf Basis von Docker Containern, Jupyter Notebook und Python Modulen.

Vorteile

- Reproduzierbare Analyse
  - Dank Docker Containern können die Daten plattformunabhängig und reproduzierbar analysiert werden.
- Flexibler Datenzugriff
  - Der Datenzugriff kann entweder vom Personal Computer, einer externen Datenbank oder direkt von einem SwisensPoleno erfolgen.
- SwisenDataExplorer integriert
  - Für schnellen Import und Export von Datensätzen aus dem Netzwerk oder Messsystem.
- Quelloffen
  - Lizenziert nach dem GLP-Standard, ermöglicht eigenständige Optimierungen und erweiterte Anwendungsbereiche.

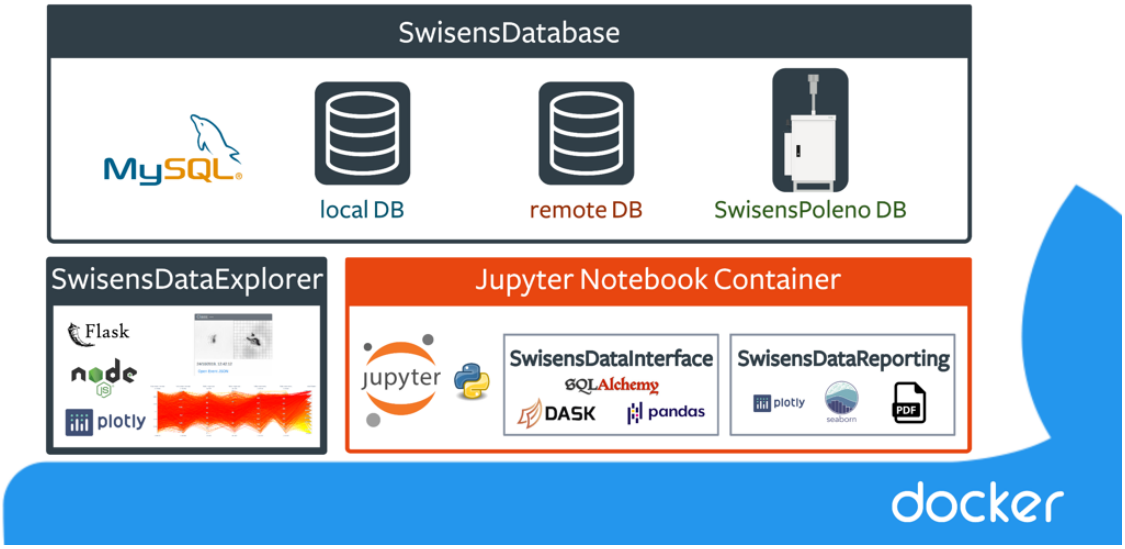


Bild 5: Übersicht SwisensDataAnalyzer

SwisensDatabase

SwisensDatabase ist eine Serversoftware zum Datenmanagement der Messinstrumente eines Messnetzes. Swisens setzte auf Open Source Standardkomponenten und offene Protokolle.

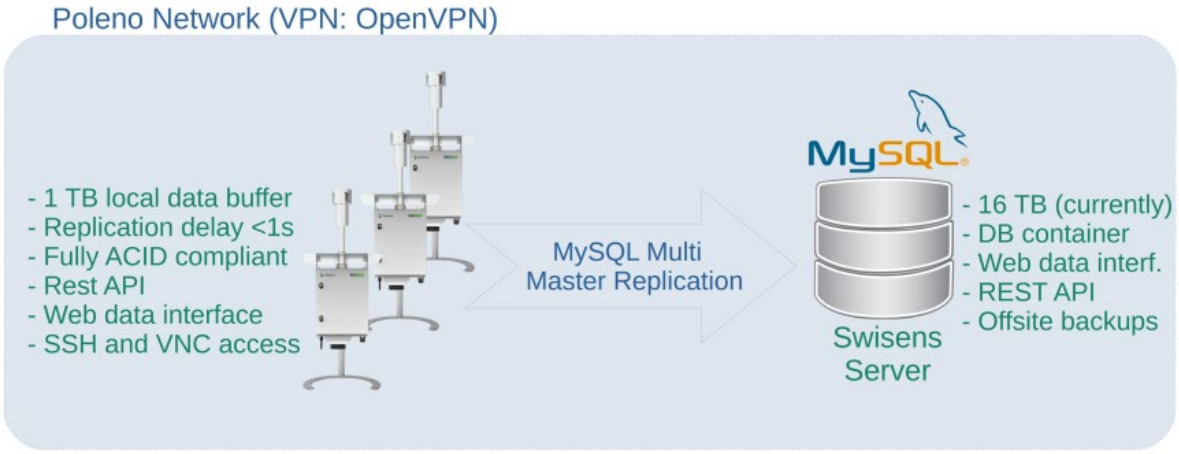


Bild 6: Übersicht SwisensData Server mit Schnittstellen

Menge der Messdaten

Aus dem Holografieverfahren liegen bis zu vier Bilder der Partikel vor. Die anderen Methoden liefern nummerische Werte. Pro gemessenes Aerosolpartikel sind das typisch vier Bilder à 80kB und nochmals 80kB Daten und somit total 400kB/Partikel.

Wenn jede Sekunde 30 Partikel gemessen werden, gibt das pro Tag mehr als 1000 MB Daten. In der Praxis gibt es zum Teil sehr hohe Konzentrationen und es gibt mehrere Gigabytes Messdaten pro Tag.

Datenformat und Metadaten

Als Datenformat der Messdaten wird JSON verwendet, die Bilder werden im verlustfreien PNG Format gespeichert und die Datensätze können mittels ZIP archiviert werden. Weiter hat JSON den Vorteil , dass es einfach lesbar für Menschen ist und auf Grund der Namen weitgehend selbsterklärend. Zum Lesen der Daten gibt es Erweiterungen für den Browser z.B. JSONView Addon für Firefox. Bild 7 zeigt eine Ausschnitt der Eventdaten vom SwisensPoleno.

Die Messdaten von einem Partikel werden als Event bezeichnet da es in diesem Datensatz nicht nur Messdaten sondern auch Resultate aus Analysen und Metadaten hat. Somit hat man hier keine Trennung zwischen Messdaten und Resultaten. Die Bild 7 zeigt den Abschnitt des Eventdatenfiles mit den Resultaten der Klassifikation und den Metadaten.

Die Variable eventBaseName ist Teil der Name des Eventfiles als auch der zugehörigen Holografiebilder. So kann die Verlinkung zwischen den Eventdaten und den Bildern gemacht werden.

Bild 7: Beispiel Messdaten im JSON Format (links) sowie Zusatz- und Metadaten in einem Messdatenfile (rechts).

Datenqualität und Einbettung von Analysedaten

Für die Überprüfung der Datenqualität werden unmittelbar nach der Messung gewisse Analysen durchgeführt und die Resultate im Eventdatensatz gespeichert. Diese Analyseresultate ermöglichen dann zum Beispiel eine Überprüfung der Qualität oder können für die Ausfilterung von ungenügenden Datensätzen verwendet werden. Bild 8 zeigt zum Beispiel, wie nummerische Analyseresultate, hier Kennwerte für die Morphologie eines Aerosolpartikels, welche von der Analyse der Holografiebilder stammen, im Eventdatensatz gespeichert werden.

Art der Datenspeicherung bzw. des Datenzugriffs: Datenfile versus Datenbank

Für das Arbeiten mit den Daten hat sich gezeigt, dass die Verwendung einer Datenbank Vorteile bringt im Vergleich zur Arbeit mit den gespeicherten Eventdaten. Es ist einfacher und mehr praxisorientiert. Zum Beispiel ist das Ausführen von Analysen bedeutend schneller, wenn eine Datenbank verwendet wird.

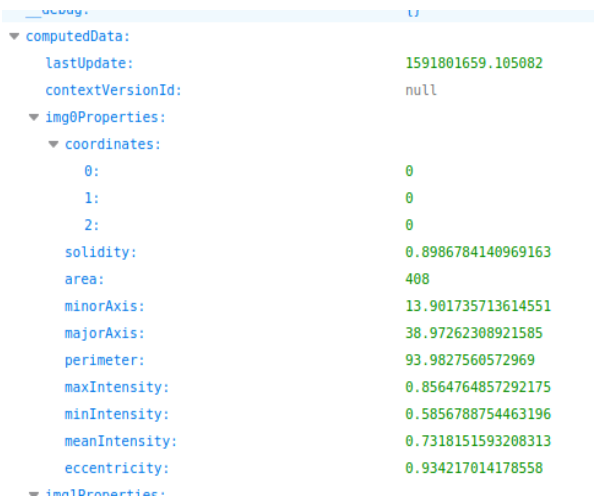


Bild 8: Resultate von Analysen im Eventdatensatz

Gewählter Typ der Datenbank

Es wird eine MySQL Datenbank verwendet. Die Hauptgründe dafür waren, dass MySQL eines der grössten und weltweit am häufigsten eingesetztes System für relationale Datenbanken ist. MySQL steht unter der GPL-Lizenz und gehört damit zum Bereich der Open Source-Software. Es gibt ausserdem eine kommerzielle Version, welche einen vollständigen Support beinhaltet. Es ist auf vielen unterschiedlichen Plattformen einsetzbar, für ein hohes Speichervolumen und bietet eine schnelle Auswertung der Abfragen. Quelle https://www.datenbanken-verstehen.de/datenbankarten/mysql/

Datenübertragungsprotokoll

Die Datenübertragung vom Messinstrument zum Nutzer kann entweder über Kabel/Ethernet, über ein GSM Modem oder über Datenspeicher gemacht werden.

Verwendetes Betriebssystem im Messinstrument

Das Messinstrument hat einen Industrie-PC eingebaut auf welchem Xubuntu (xubuntu.org) als Betriebssystem läuft. Der Vorteil ist, dass es sich um eine freie und kostenlose Linux-Distribution, die von Debian abstammt. Xubuntu verwendet die Xfce-Desktop-Umgebung, die vor allem für ihre Stabilität und ressourcenschonende Arbeitsweise bekannt ist. (https://de.wikipedia.org/wiki/Xubuntu)

Eingesetzte Software für die Datenverarbeitung

Als Programmiersprache verwendet Swisens Python (https://www.python.org/). Die vielfältigen Bibliotheken die Plattformunabhängigkeit und die Popularität sprechen für sich.

Bild 9: Übersicht eingesetzte Software für die Swisens Lösung

Bild 9 listet auf für was welche Software beim SwisensDataExplorer und beim SwisensDataAnalyser eingesetzt wird. Als Webframework wird das Python basierte **Flask** verwendet (https://palletsprojects.com/p/flask/). Weiter wird **Node.js**, eine plattformübergreifende Open-Source-JavaScript-Laufzeitumgebung, die JavaScript-Code außerhalb eines Webbrowsers ausführen kann eingesetzt (https://nodejs.org). **Jupyter Notebooks** wird verwendet so dass ein interaktives Arbeiten mit den BigData einfach ermöglicht wird und für die Erweiterung des Codes durch Dritte. Das ist sozusagen die BigData Alternative für Excel (https://jupyter.org/). **DASK** wird eingesetzt um die Verarbeitung von BigData mit einfachen Computern zu ermöglichen. Dask erreicht dies durch Parallelisierung (https://docs.dask.org/en/latest/). **Plotly** (https://plotly.com/) wird verwendet um interaktive, web-basierte Grafiken zur Datenvisualisierung zu erstellen.

Integration von künstlicher Intelligenz

Bild 10 zeigt das die im SwisensDataExplorer verwendeten Softwaretools für die Verwendung von künstlicher Intelligenz. Zusätzlich zu den oben erwähnten Tools kommt TensorFlow hinzu (https://www.tensorflow.org/). TensorFlow ist eine Open Source-Plattform für maschinelles lernen. Damit können Benutzer ohne spezielles Know-how zu Machine Learning die Algorithmen für den SwisensPoleno erweitern und so ihr System Ihren Bedürfnissen anpassen. Zum Beispiel bei der Echtzeit Pollenidentifikation. Ein Betreiber kann die Erkennungsfähigkeit auf spezifische regionale Pflanzen welche Pollen emittieren erweitern.

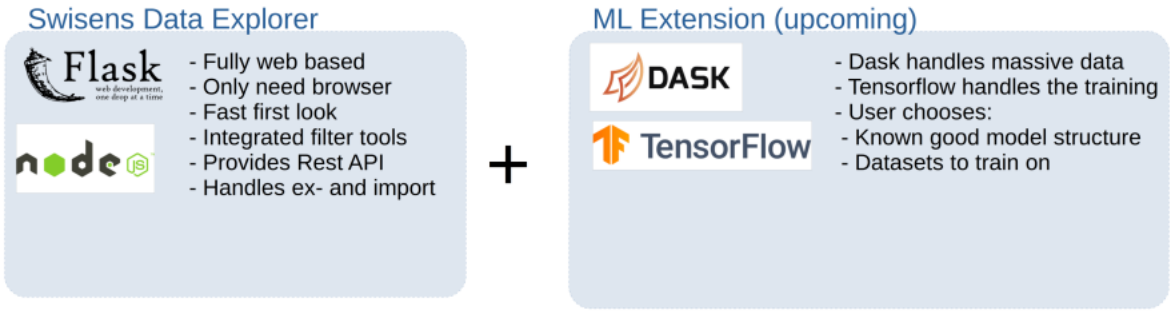


Bild 10: Machine Learning Toolset

6.1.2 Weiteres

Verfügbarkeit des Codes

Der SwisensDataAnalyser steht unter GPL. Der Source Code steht auf Anfrage zur Verfügung. (E-Mail an erny.niederberger@swisens.ch)

Anforderungen

Die Anforderungen sind im Dokument Anforderungsliste_BIG_DATA_ITC_BDC_PRJ_01.xlsx ersichtlich, welches auch im Anhang (Kapitel 7.3) zu finden ist. Bemerkenswert ist, dass typischerweise 3 Gigabyte Messdaten pro Tag von einem Messinstrument produziert werden. Eine Zeitserie über eine typische Messperiode von neun Monaten hat einen Umfang von bis zu 1 Terabyte. Die Messdaten werden mit Maschine Learning Algorithmen bearbeitet und als Resultate liegen die Klassifizierungen für verschiedene Pollensorten vor.

Präsentation zu den von Swisens verwendeten Tools

Eine Übersicht der Lösung für das Bewältigen dieser Datenmengen gibt die Präsentation, die anlässlich des Forschungsplenum der HSLU T&A vom 25.11.2020 von Yanick Zeder von Swisens gehalten wurde. Die Präsentation ist im Anhang (Kapitel 7.4) zu finden.

6.2 Sensor Data Cloud

Die «Sensor Data Cloud» wurde an der HSLU entwickelt, um die Datenerfassung, -speicherung, -visualisierung und Weiterverarbeitung von Sensordaten zu vereinfachen und auch zu automatisieren. Die Lösung ist so aufgebaut, dass sie in verschiedenen Projekten anwendbar ist und auch von weniger IT-affinen Anwendern mit wenig Zeitaufwand installiert und eingesetzt werden kann. Die Hauptentwicklungsarbeit wurde dabei vom Institut für Elektrotechnik IET durchgeführt. Die «Sensor Data Cloud» wurde dann auch in drei konkreten Projekten eingesetzt, die im nachfolgenden Kapitel 6.3 beschrieben sind.

Die Sensor Data Cloud besteht aus den folgenden drei Modulen die nachfolgend beschrieben werden:

- Mini Data Cloud (MDC)
- Python-MDC-Client
- Python-DAQ-Example

Die detaillierten Programmcodes inklusive Installationsdateien und -anweisungen sind im GitLab des EnterpriseLabs der HSLU für alle Personen, die einen HSLU Login besitzen, frei zugänglich (https://gitlab.enterpriselab.ch/sensor-data-cloud). Die einzige Voraussetzung für den Zugriff ist eine einmalige Registrierung für das EnterpriseLab (https://eportal.enterpriselab.ch/#/registration)

6.2.1 Mini Data Cloud (MDC)

Die MDC ist der Hauptteil der Sensor Data Cloud mit dem eine sichere Datenerfassung in dezentralen Systemen ermöglicht wird. Die weitere Dokumentation und die Programmcodes sind hier zu finden: https://gitlab.enterpriselab.ch/sensor-data-cloud/mdc. Die Lösung beinhaltet folgende Möglichkeiten und Eigenschaften:

- Vor-konfiguriertes System für die Einrichtung und die Durchführung einer Testdatenerfassung (DAQ) in kürzester Zeit von Grund auf.
- Modularer Aufbau mit Docker-Containern zur einfachen Anpassung.
- Läuft auf jedem Linux-System mit einer statischen IP, mit jeder beliebigen virtuellen Maschine- mit verschiedenen Anforderungen (z. B. SNF-Richtlinien).
    - Der Einsatz von SwitchEngines für das Einrichten einer virtuellen Maschine wird empfohlen.
    - Alternativ kann auch ein lokales NAS verwendet werden.
- Sichere Datenübertragung mit SSL-Verschlüsselung im gesamten System.
    - Ein Aufbau ohne SSL-Verschlüsselung ist ebenfalls möglich für schnelle Tests in lokalen Netzwerken.
- Aufruf von automatisierten Python3-Jobs nach Intervall oder einer Tageszeit.

Die Mini-Data-Cloud besteht aus einzelnen Modulen/Containern und kann individuell angepasst werden, um Betriebs- und Sensordaten zu erfassen. Die Architektur der MDC ist in Bild 11 zu sehen und besteht aus den folgenden Modulen:

MQTT

Zur Datenübertragung wird das MQTT (Message Queuing Telemetry Transport) Protokoll eingesetzt MQTT ist ein offenes Netzwerkprotokoll für Machine-to-Machine-Kommunikation, das die Übertragung von Telemetriedaten in Form von Nachrichten zwischen Geräten ermöglicht, trotz hoher Verzögerungen oder beschränkter Netzwerke. (Quelle: www.wikipedia.org)

Mosquito

Ist der eingesetzte MQTT-Broker. Ein MQTT Broker steht im Mittelpunkt jedes Publish / Subscribe-Protokolls und kann zahlreiche gleichzeitig verbundene MQTT Client verwalten. Der Broker ist dafür verantwortlich, alle Nachrichten zu empfangen, die Nachrichten zu filtern, zu bestimmen, wer die einzelnen Nachrichten abonniert hat, und die Nachricht an diese abonnierten Clients zu senden. Sollte die Verbindung von einem abonnierenden Client zu einem Broker unterbrochen werden, puffert der Broker die Nachrichten und sendet sie an den Abonnenten, wenn dieser wieder online ist. Wenn die Verbindung vom Publishing-Client zum Broker ohne Benachrichtigung getrennt wird, kann der Broker die Verbindung trennen und den Abonnenten eine zwischengespeicherte Nachricht mit Anweisungen des Publishers senden. (Quelle: www.wikipedia.org, https://www.opc-router.de/was-ist-mqtt/)

Node-Red

Wird zur Verbindung von Hardware, Schnittstellen und Services eingesetzt. Node-RED nutzt dabei eine grafische Programmierung und nutzt dabei sogenannte **nodes (Knoten)**, welche zu einem **flow** verbunden werden. Jeder Node hat einen klar definierten Zweck; ihm werden Daten gegeben, er macht etwas mit diesen Daten und gibt diese Daten dann weiter. Das Netzwerk ist für den Datenfluss zwischen den Knoten verantwortlich. (Quelle: https://nodered.org)

Nginx

Server, der üblicherweise als HTTP Server, Reverse Proxy Server und Mail Proxy Server eingesetzt wird. Der Webserver nginx zeichnet sich besonders durch seinen geringen Bedarf an Hauptspeicher und CPU bei vielen gleichzeitigen Verbindungen aus. (Quelle: https://www.nginx.com/)

Lets Encrypt

Let's Encrypt ist eine freie, automatisierte und offene Zertifizierungsstelle, die digitale Zertifikate, die sie zur Aktivierung von HTTPS (SSL/TLS) auf Webseiten benötigt werden, kostenlos zur Verfügung stellen. (Quelle: https://letsencrypt.org/de/how-it-works/)

Influx DB

Datenbank zur Datenspeicherung, die für Zeitreihen-Daten optimiert ist.

Grafana

Grafana ist eine plattformübergreifende Open-Source-Anwendung zur grafischen Darstellung von Daten aus verschiedenen Datenquellen wie z. B. InfluxDB, MySQL, PostgreSQL. Die erfassten Rohdaten lassen sich anschließend in verschiedenen Anzeigeformen ausgeben. Einfache Operationen (z.B. Multiplikation, einfache Datenfilter) können ebenfalls mit Grafana durchgeführt werden. (Quelle: www.wikipedia.org)

Cronjobs

Zum Ausführen von automatischen wiederkehrenden Aufgaben, Daten-Weiterverarbeitung und Analysen, die komplexer sind als dass sie mit Grafana durchgeführt werden können und mit Python programmiert werden. Alternativ könnte hier sicherlich auch ein Workflow-Management-Tool wie Apache Airflow (https://airflow.apache.org/) eingesetzt werden, welches bei Bedarf separat installiert werden müsste. Siehe dazu https://gitlab.enterpriselab.ch/sensor-data-cloud/apache-airflow.

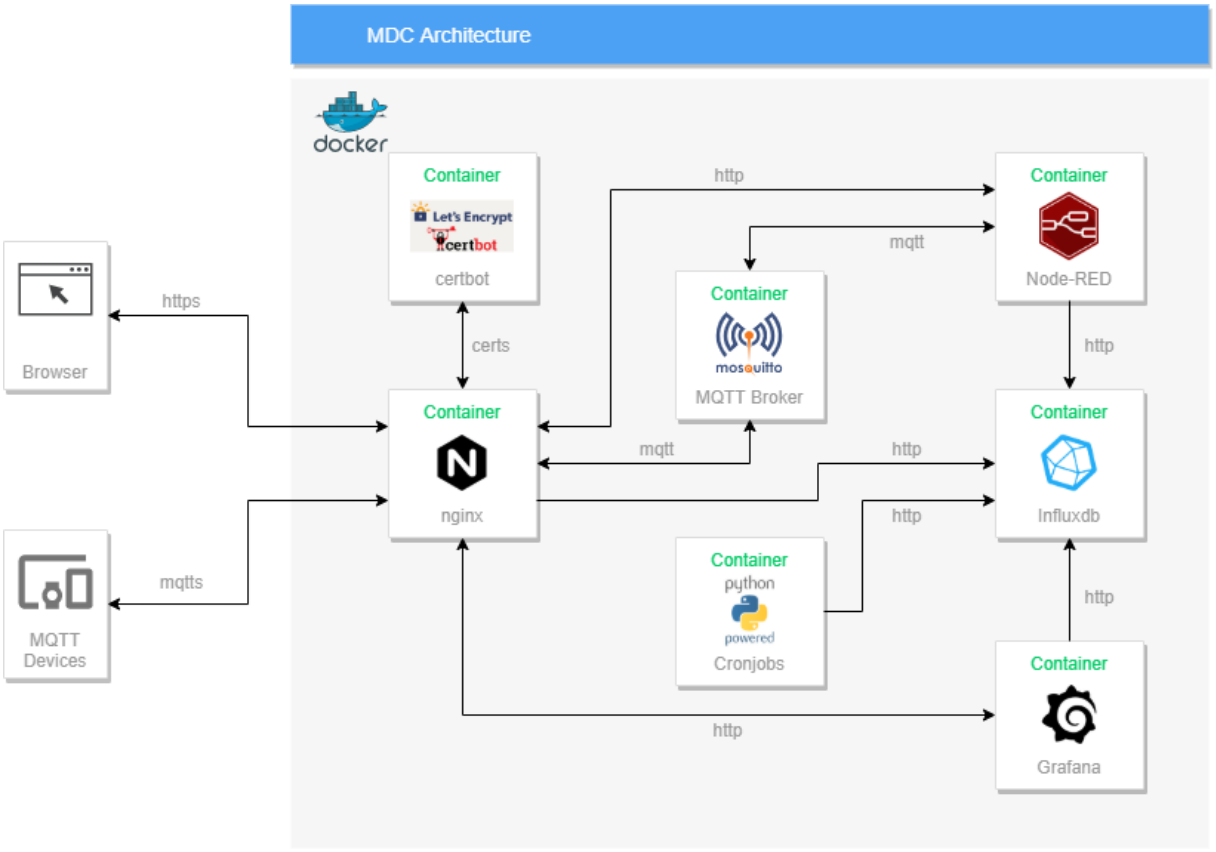


Bild 11: Architektur der MDC

6.2.2 Python-MDC-Client

Dieses Python3 Modul dient dem Zugriff auf die Influx Datenbank, um die Daten nachzubearbeiten und Analysen durchzuführen. Dabei werden die Daten aus der Influx DB in ein Pandas DataFrame umgewandelt. Das Python MDC Modul hat folgende Funktionalitäten:

- Listet alle verfügbaren Datensätze auf
- Kann Datensätze lesen
- Gibt zeitlich gemittelte Werte aus für große Datensätze
- Kann zusätzliche Daten an Datensätze in der InfluxDB anhängen (z. B. berechnete Signale aus der Nachbearbeitung)
- Kann Datensätze löschen
- Langzeit Backup im HDF5 Datenformat

Der Python-MDC-Client basiert auf https://influxdb-python.readthedocs.io und beinhaltet auch die Möglichkeit zur Umwandlung des Zeitstempels von UTC.

6.2.3 Python-DAQ-Example

Python3-Beispiel zum Senden von Testdaten an die Mini Data Cloud. Dieses Beispiel sendet gruppierte Zeitstempel als ZIP-komprimierte CSV-Dateien, um Overhead und Datenverbrauch in mobilen Netzwerken (z. B. LTE) zu reduzieren. Dies macht die Datenübertragung auch zuverlässiger, insbesondere wenn die Internetverbindung von Zeit zu Zeit unterbrochen werden kann. Zusätzlich wird in diesem Modul der Sensor Data Cloud auch die Konfiguration des Rasbperry PIs erklärt sowie Tipps und Hinweise zum LTE Mobilnetz inklusive Datenkomprimierung gegeben.

6.3 Projektbeispiele mit Nutzung der Sensor Data Cloud

6.3.1 Autonome Low-Cost Emissionsüberwachung von Holzfeuerungen

Projekt Idee

Holzfeuerungen sind CO2-neutrale Energielieferanten mit erheblichem Ausbaupotenzial und daher ein wichtiger Bestandteil der schweizerischen Energiestrategie 2050. Allerdings ist die Holzverbrennung für einen signifikanten Anteil an der Luftverschmutzung verantwortlich. Zur Überprüfung der Einhaltung der LRV-Grenzwerte von automatischen Holzfeuerungen > 70 kW finden periodische Emissionsmessungen vor Ort statt, die in der Regel weniger als 2 Stunden dauern und zudem nur den stationären Betrieb erfassen. Da im Praxisbetrieb aber häufig auch instabile Betriebszustände wie An- und Abfahren vorkommen und der Betrieb oft nur für die Emissionsmessung optimiert wird, können die realen Emissionen deutlich höher sein.

Bei schlecht betriebenen Anlagen oder bei Anlagen, die sich in oder nahe an dicht besiedelten Gebieten befinden, kommt es häufig zu Klagefällen durch die Anrainer. Zusätzlich werden immer mehr Anlagen in Massnahmengebieten gebaut. Daher haben die Behörden ein erhöhtes Interesse zur Überwachung der Emissionen von Holzfeuerungsanlagen. Jedoch sind die finanziellen und personellen Ressourcen bei den kantonalen Vollzugsbehörden für Emissionsmessungen knapp. Falls dann Emissionsmessungen über längere Zeiträume durchgeführt werden, dauern diese in der Regel nur zwischen 3 bis 5 Tage.

Die Holzfeuerungsbetreiber haben ihrerseits mit immer strengeren Auflagen zu kämpfen, die sich negativ auf die Wirtschaftlichkeit auswirken. Zum Beispiel wird immer häufiger auch für kleinere und mittlere Anlagengrössen (500 kW – 5 MW) eine kontinuierliche Emissionsmessung vorgeschrieben, die sehr teuer ist (>60'000 CHF plus jährliche Wartungskosten von mehreren tausend bis 10'000 CHF).

Hohe Emissionen von schlecht betriebenen Holzfeuerungsanlagen könnten in den meisten Fällen durch eine Betriebsoptimierung reduziert werden und in vielen Fällen dadurch gleichzeitig der oft niedrige Wirkungsgrad erhöht werden. Die auf den Anlagen verfügbaren Daten sind dazu in der Regel jedoch nicht ausreichend. Da Messungen von Emissionen mit Standardmessgeräten und die Erfassung wichtiger Anlageparameter über einen längeren Zeitraum, wie für eine Betriebsoptimierung notwendig, teuer sind, werden solche Langzeitmessungen in der Regel nur bei Klagefällen und zudem nur während 3 bis 5 Tagen durchgeführt.

Daher wurde in diesem Projekt ein günstiges Messsystem (< 3500 CHF) entwickelt und im Labor getestet, mit dem wartungsarme und autonome Langzeitmessungen (LZM) von mindestens 3 Wochen durchgeführt werden können. Bei Messdauern, die länger als 3 Wochen dauern, oder bei häufigeren Messungen auf der gleichen Anlage besteht zusätzlich zur Betriebsoptimierung des Feuerungsbetriebs die Möglichkeit zur Früherkennung von Schäden an der Anlage und zur Optimierung von Wartungseinsätzen

Realisierung

Es wurde ein automatisiertes Messsystem entwickelt, das eine autonome Überprüfung der Sensoren mit Plausibilitätstests und Alarmierung bei Fehlfunktion, eine verschlüsselte Datenübertragung in die Sensor Data Cloud eine zentrale Datenbank und eine Quasi-Echtzeit-Visualisierung der Daten und Auswertung im Internet beinhaltet. Das Messsystem erfasst die wichtigsten gasförmigen Emissionen (CO und NO_x) und Parameter des Betriebs der Holzfeuerung (O_2, Abgastemperatur, Betrieb des Abgas- oder Primärluftventilators) und des Elektroabscheiders. Ein Prinzipschema ist in Bild 12 zu finden.

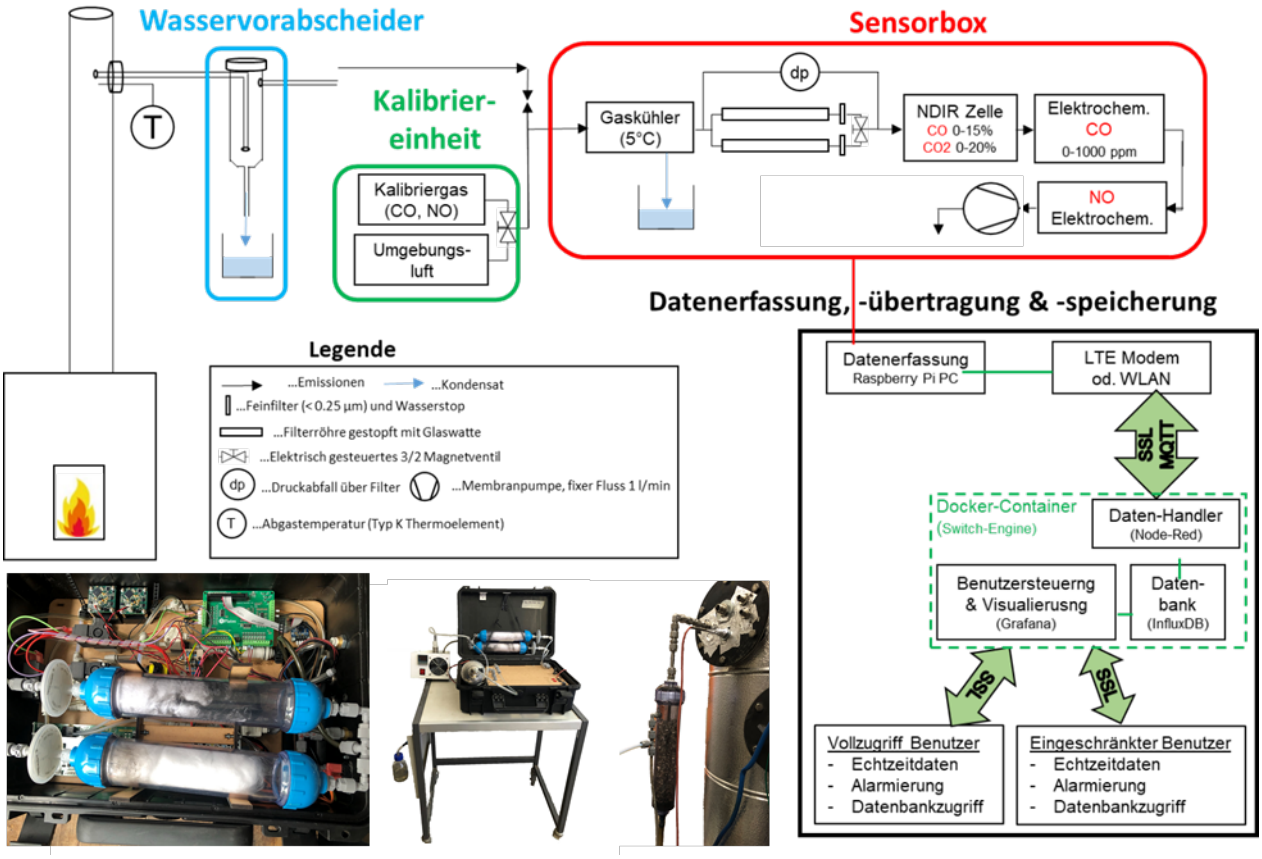


Bild 12: Schematische Darstellung des Messsystems.

Die Messtechnikhardware besteht ausschliesslich aus kommerziell verfügbaren Komponenten. Es besteht aus einem NDIR Gassensor zur Messung von CO und CO_2, sowie elektrochemischen Gassensoren für den tiefen CO Messbereich und NO. Ausserdem wurde das Messsystem mit einer günstigen, einfachen und wartungsarmen Messgasaufbereitung konzipiert. Dies wurde mittels geringen Volumenströmen (1 l/min Absaugrate), den Verzicht auf die sonst üblichen beheizten Messgasleitungen bis zum Gaskühler und die automatisch umschaltbaren Filterlinien, erreicht. Zusätzlich verfügt das Messsystem über einen Anschluss für Prüfgase, mit denen automatisiert in regelmässigen Abständen, die Gassensoren überprüft werden. Das Messsystem wird mittels einem kleinen integrierten PC gesteuert, der auch die Messdaten aufgezeichnet.

Für die Datenübertragung, -speicherung, -visualisierung und –Auswertung wurde die oben beschriebene Cloud Lösung (Sensor Data Cloud,) eingesetzt. Die Daten werden von der Datenerfassung mittels LTE-Modem in die Datenbank übertragen. Zur Datenerfassung und zur Steuerung des Messsystems wird ein Raspberry Pi (RPi) mit Linux und Python3 Programmcodes verwendet.

Evaluation des Messsystems:

- Sensoren verschiedener Hersteller wurden getestet und evaluiert
  - 4 NDIR Zellen von 4 verschiedenen Herstellern
  - Je 4 elektrochemische Zellen für CO & NO von 2 verschiedenen Herstellern
- Referenzmessungen mit Standard-Emissionsmessgerät Ultramat von Siemens
- Zahlreiche Testmessungen im Labor an verschiedenen Feuerungen (siehe Tabelle 2)
- Messung mit abschliessendem Setup:
  - 3 Wochen ohne Unterbrüche im Dezember 2019 an der 150 kW Vorschubrostfeuerung im Labor der HLSU
  - Sehr gute Übereinstimmung mit dem Standardmessgerät Siemens Ultramat
    - Korrelationskoeffizient für O_2 und CO zw. 0.95 – 0.99 und für NO 0.84
    - Steigung (ausser für den CO Messbereich > 1000 ppm) im Bereich von ± 3 % von der 1:1 Linie
    - Sehr geringer Achsenabschnitt

Tabelle 2: Übersicht über die durchgeführten Test mit Feuerungen

Feuerung	Nennleistung	Anzahl Tests	Betriebszustände
Pelletofen	6 kW	10 Tage, 2 -10 h pro Tag	Voll- und Teillast
Stückholz-kessel	30 kW	6 Tage, zahlreiche Abbrände	Ganzer Abbrand
Stückholz-ofen	6 kW	8 Tage, zahlreiche Abbrände	Ganzer Abbrand
Automat. Hackschnitzelfeuerung	150 kW im HSLU Labor	2 Tests Messdauer 1-3 Wochen	Alle Betriebsphase inkl. Start & Stopp
	530 kW Praxisanlage	1 Messung über 6 Wochen	

6.3.2 Beflaggungstechnik 4.0

Projekt Idee

Die Grundidee für die Beflaggungstechnik 4.0 ist eine Vernetzung des einzelnen Fahnenmastes mit dem Hersteller ALUART über eine Cloud-Lösung und mit dem Endkunden über eine App. Die Firma ALUART ist ein Hersteller von Fahnenmasten aus Aluminium.

Mit Hilfe einfacher Sensoren werden Belastungen am Mast erfasst (Monitoring). Die Daten liefern dem Hersteller Informationen über die Beanspruchungen der Fahnenmaste (Industrie 4.0) und können dazu verwendet werden, Sturmwarnungen zu verschicken, so dass die Fahne rechtzeitig eingezogen und dadurch Schäden verhindert werden.

Realisierung

Die erfassten Daten auf dem Fahnenmast sind primär Winddaten sowie ev. zusätzliche Wetterdaten (Temperatur, Feuchtigkeit). Es wird bereits auf dem Mast eine Datenreduktion vorgenommen, da z.B. nur die maximal gemessene Windgeschwindigkeit alle 10 Sekunden verschickt wird, die Abtastrate aber wesentlich höher liegt (5 Hz). Je nach verfügbarer Energie (z.B. im Winter) kann die Abtastrate und auch die Übertragungsrate reduziert werden. Die zu übertragenden Datenmengen sind deshalb sehr gering, typischerweise ca. 50 Byte / Minute. Die Herausforderungen liegen bei der drahtlosen und energieeffizienten Übermittlung über grosse Distanzen (10 km), weshalb ein LoRa Netzwerk (Long Range Wide Area Network) verwendet wird.

Der Fokus im Projekt liegt bei der Entwicklung eines Prototyps für eine «energieautarke Wetterstation» auf einem Fahnenmast und der Realisierung der Datenübertragung mittels LoRa. Bei grossen Distanzen verlängert sich die Sendezeit was dazu führen kann, dass weniger häufig Daten übertragen werden dürfen. Dabei handelt es sich um regulatorische Gründe sowie der Fair-Use-Policy von The Things Network, dessen LoRaWAN Netzwerk im Projekt verwendet wird. Als Energiequelle werden Solarpanels und als Energiespeicher Supercaps eingesetzt. Der Prototyp ist beliebig ausbaubar, z.B. in der Gebäudeautomation (Storen vor Unwetter schützen, nachhaltiges, energiesparendes Beschattungssystem). Zur Datenübertagung, -speicherung und -verarbeitung wurde eine frühe Version der MDC, die ebenfalls vom IEE implementiert wurde, angewendet.

Verwendete Tools - Übersicht

Tool	Beschreibung
LoRaWAN	Datenübertragungsprotokoll von Fahnenmast 4.0 zur Zentralen Recheneinheit
InfluxDB	Datenbank zur Datenspeicherung
C, JavaScript	Programmiersprachen zur Programmierung der Mikrokontroller in Fahnenmast 4.0 (C) und der Zentralen Cloud Lösung (JavaScript)
Docker, Mosquitto, NodeRed	Tools der IoT Cloud Lösung zur Datenübertragung, und -verabeitung.
Grafana	Daten Visualisierungspackages

6.3.3 Monitoring DB IGE

Projekt Idee

Die Monitoring DB ist quasi der Vorgänger der Mini Data Cloud (MDC) und beinhaltet die gleichen Grundbausteine wie die InfluxDB und Grafana. Es wurde eine Datenbankinfrastruktur zur Verfügung gestellt, welche den IGE Data Scientist einen einfachen und schnellen Zugang zu den Messdaten gewährt. Das System bewährte sich in einem grösseren Projekt, wo schlussendlich über 260 Mio. Zeitstempel/Wertepaare abgespeichert wurden und über die Statistiksoftware R für Analysen abgefragt wurden. Kritisch zu erwähnen bleibt der Unterhalt und die Wartung des Systems. Dies bedarf Personal mit entsprechendem Know-How und Kompetenz sowie Projektbudget, welches nach Projektabschluss nicht mehr zur Verfügung steht. Zudem erwiesen sich die Themen Benutzerrechte auf dem Monitoring System und Datensicherheit als komplex und der aktuelle Prototyp weist diesbezüglich sicherlich noch Mängel auf.

Realisierung

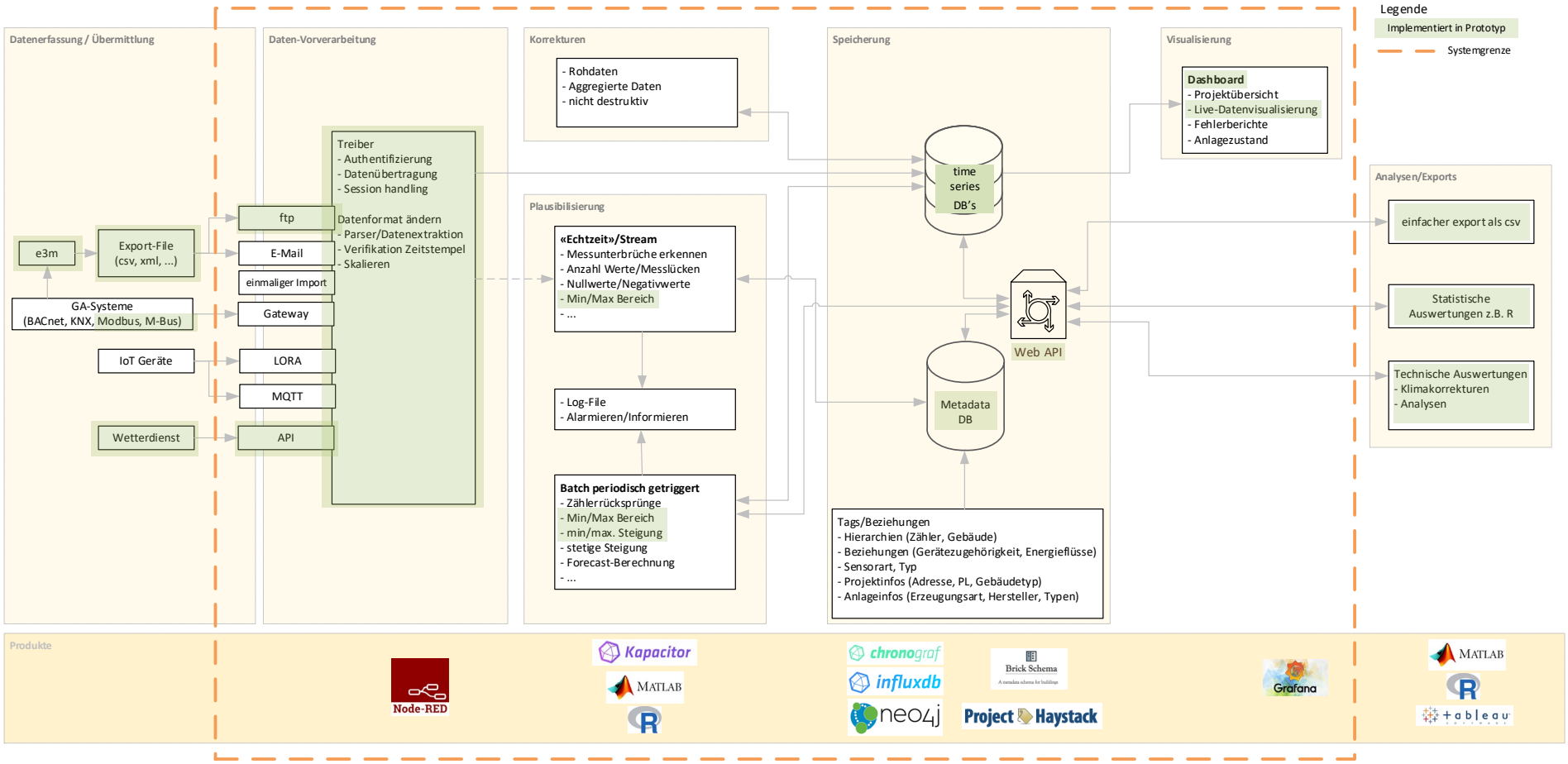


Bild 13: Softwarearchitektur der MonitoringDB IGE

7 Anhang

7.1 Interview zum Forschungsdatenmanagement

Interview von Michael Wanner Projektleiter des Projekts «Forschungsdatenmanagement – Umsetzung» mit Erny Niederberger HSLU T&A IET, 30.6.2020

Die spezifischen Antworten beziehen sich auf das APID Projekt des IET und den daraus herausgegangenen Start-up Swisens AG

Frage sind in blauer Schrift und Antworten in kursiver schwarzer Schrift ausgeführt.

- Welche Daten werden erhoben?
  - Format z.B. Video, Audio, Messungen, Text etc.?
    APID Projekt: Messdaten und Metadaten: Nummerische Daten in JSON Format und Bilddateien und Analysedaten die im Event-datensatz gespeichert sind.
  - Welche Daten-Management-Tools verwenden sie?
    Typ der Datenbank: MySQL Datenbank
    Synchronisationstools: => Integrierte Replikation von MySQL,
    Backup: Redundante Disk, bzw. redundantes Serversystem, skriptbasierte Automatisierung
    Datenanalyse und Bearbeitung mit Python:
    Datenbank: Zugriff von Kunden bzw. externen über Rest-API
    Interner Zugriff: für Skripte die auf dem Server laufen, Zugriff direkt mit SQL aus python heraus
    Datenvisualisierung, Analyse und Management: Swisens Data-Explorer Software: Rest-API Back-end: FLASK-Library von Python, Java-Script
    Entwicklung: Datenanalyse, mit Jupiter Notebook:
    Daten teilen mit dritten. Z.B. Forscher
    Server-Software die auf unserem Server läuft: Produkt: Pydio Cells Home Edition und Switch-drive
  - Welche Art Daten werden erhoben? Unpersönliche, persönliche, sensible, vertrauliche, geheime, fach- oder gerätespezifische Daten (ethische, rechtliche Abklärungen etc.)?
    Unpersönliche, gerätespezifische Daten

- Gibt es Richtlinien oder Vorgaben, die Ihren Umgang mit FD genauer bestimmen (z.B. auch von Partnern, IP, Urheberrecht, Patente etc.)?
  Nein
  - Analyse: Erfolgt eine Anonymisierung? Wenn ja, wie?
    Nein, erfolgt nicht.
- Werden heute Daten intern/extern geteilt/ausgetauscht (Kollaboration)?
  Ja, intern und extern
  - Wie werden sie ausgetauscht (Mail etc.)?
    Intern: *Wir betreiben einen eigenen Server für die SQL Datenbank*
    Extern: Zusätzlich ein weiterer Dienst für das Teilen von Messdaten. SW: Pydio Cells Home Edition. Damit können wir selektiv die Daten teilen. Diese sind jedoch nicht öffentlich sondern der Zugriff erfolgt über ein account management
  - Regelung der Zugriffsrechte?
    Via Login zur Datenbank
- Wie werden sie heute dokumentiert (Versionskontrolle)?
  Jeder Messevent erhält einen eindeutigen Bezeichner
  Im Messdatensatz ist der Filename integriert für den Fall, wenn der Eventdatensatz als File gespeichert wird. Zusätzlich wird der Zeitpunkt wenn das File erzeugt worden ist im Datensatz gespeichert. So kann der Zeitpunkt wann die letzte Analyse gemacht wurde nachverfolgt werden.
  Ein Messevent enthält neben den Rohdaten noch Metadaten und Analyseresultate: Für die Nachvollziehbarkeit der Analyseresultate wird die Version der Berechnungssoftware gespeichert
  Bsp:
  Event-ID: Unique Identifier
  EventBaseName: Filename wenn die Daten als File abgespeichert werden

utcEvent: Moment wenn die Messung erfolgt ist
utcJson: Wenn das File geschrieben worden ist (Versionierung)
Version des Codes mit welchem die Auswertung gemacht worden ist

- Wie und wo werden sie heute aufbewahrt/gespeichert (Backup)?
  Im Moment werden die Daten auf einem Server gelagert. Der Server hat eine Raid konfiguration. Backup machen wir selber.
- Wie und wo werden sie archiviert (heute) und wie lange?
  Das ist noch nicht gelöst. Eine Zeitserie der Messevents von einem Jahr und Messystem hat die Grösse von ca. 2Terabyte. Wir werden eine Publikation machen dieses Jahr und müssen noch eine Lösung finden.
  - Können die Daten wiederverwertet werden?
    Ja
  - Welche Massnahmen treffen Sie, damit die Daten nachvollziehbar bleiben (Metadaten für die Beschreibung der Daten etc.)?
    Für jedes Messereignis wird ein Messevent erzeugt. Metadaten eines Events: Eventidentifikations ID, Typ des Messsystems, Seriennummer des Messsystems, Version des Messsystems, EventBaseName, Ort an welchem das Messystem steht inklusive Höhe über Meer, Zeit des Messevents.
    Für Teil der Analyseresultate: Version des Codes für die Analyse
  - Wie kann man sie finden (Publikation, Plattform, Journal, Repository)?
    Bis jetzt noch nicht public
  - Was für ein Datenarchiv würden Sie dafür am ehesten nutzen? Z.B Archiv des Instituts, zentrales lokales Archiv der HSLU, nationales fachspezifisches Archiv?
    Es gibt kein nationales fachspezifisches Archiv
    Dann dies der HSLU
    Am liebsten würde ich eines nutzen in welche die Forschungsdaten zusammen mit der Datenbank und den Analysetools als Container zur Verfügung gestellt werden. Damit könnten später jederzeit jedermann relativ einfach mit den Daten arbeiten. Das entspricht am ehesten dem was mit den FAIR Data Prinzipien gewünscht wird und ist so technisch relativ einfach zu realisieren.
- Werden Daten gelöscht? Wenn ja, welche?
  Ja: Die Rohdaten der Kamerabilder mit hoher Auflösung werden nicht gespeichert. Nur das Resultat der holografische Rekunstruktion wird als Bild gespeichert mit einer Auflösung von 200x200 Pixel als Grauwertbild. Diese «ausschnitt-» Bilder werden dann weiter verarbeitet.

- *Haben sie schon einmal einen DPM ausgefüllt?*
  Ja
  - *Welche Punkte daraus waren am schwierigsten zu definieren?*
    Datenerhalt, Planung des Datenerhalts, wie und wo werden die Daten zugänglich gemacht, Interpretation der Punkte zum FAIR data principle nicht klar.
    Reusability: Zum Beispiel: Sind Terabyte von Daten auf einem Bandlaufwerk gespeichert wirklich als zugreifbar klassifizierbar. Theoretisch ja, in der Praxis ist der Aufwand schon sehr gross und die accessibility ist als sehr schwierig zu bewerten. Reicht das aus?
    Meiner Meinung nach kann das nicht das Ziel sein. Das Ziel ist eine einfache reusability. Das bedeutet, dass die Daten zugriffsfreundlich gelagert werden.
  - *Welches sind die grössten Herausforderungen bezüglich Datenmanagement?*
    Dass vieles unklar ist
    Es gibt wenig Erfahrung ,vor allem wenn man grössere Datenmengen hat (Kompetente Beratung mit hoher Fachexpertise zu Big Data fehlt)
    Wir konnten niemanden finden der aus einem grossen Erfahrungsschatz zurück greifen konnte.
    Dass es um so schwieriger wird die Daten «umzuzügeln» je grösser die Datenmengen sind. Der Aufwand für nachträgliche Aufdatierung mit Metadaten etc. wird immer grösser
    Die grosse Menge, die Verarbeitung innert nützlicher Frist, die benötigte Hardware zur Bearbeitung (RAM), Nähe von Daten zur Verarbeitung,
    Methoden: Wahl der Datenbank im Bezug auf das gemischte Format der Daten
  - *Übersicht haben über die Angebote von externen Dienstleistern (Amazon, IBM, Microsoft Azure etc.) und vor allem über deren Nachteile und Kosten:*
    Erfahrungsaustausch dazu notwendig => Schwierig an Personen heran zu kommen die Kompetent sind, neutral und in der Lage das verständlich zu kommunizieren.
  - Wo haben sie am meisten Probleme beim Eingabe-Prozess? Was genau sind die Probleme?
    Diese Frage verstehe ich nicht
  - Holen sie Hilfe und wenn ja bei wem? Wenn nein, warum nicht?
    Bibliothek
    Enhancer Team (ETHZ)

- Welche zentralen Serviceleistungen (Beratung und Schulung) rund um Forschungsdaten würden Sie sich von der HSLU wünschen?

  Dringend notwendig ist ein Team von erfahrenen Personen (Scientific-IT-Team) welche kompetent zu folgenden Themen Support leisten kann und Schulungen anbieten kann:

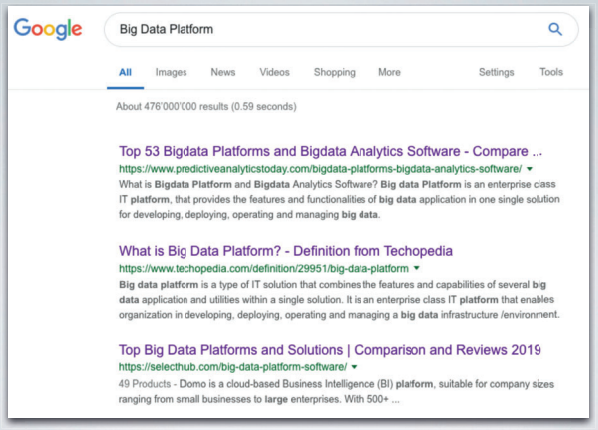
  - *Zur Verfügungsstellung von Infrastruktur für die Daten und Tool-Empfehlungen, so dass von Anfang an für die ganze Lebenszeit der Daten im Projekt und darüber hinaus mit der der selben Plattform gearbeitet werden kann, so dass kein extra Aufwand für die zur Verfügungsstellung für weitere Forscher oder eine Publikation entsteht.*
  - *Integrierte Back-up Lösung so dass nicht jeder selber eine Backup-Lösung machen muss (Problem Switchdrive => kein garantiertes Backup)*
  - *Beratung bei der Auswahl des Typs der Datenbank im Kontext der Art der Daten*
  - *Beratung über Best-practice von anderen Projekten, Fachhochschulen, Universitäten und Hochschulen zum Verhindern dass jeder das Rad neu erfindet*
  - *Beratung über Serviceangebote zum Datenmanagment mit Big-Data, von kommerziellen Anbietern*
  - *Schulung zu Datenmanagement, Big-Data Analyse, Analysetools, Visualisierungstools*
  - *Beratung über rechtliche und ethische Fragen zu Zugangsbeschränkungen, Umgang mit sensiblen Daten, Datenschutz*
  - *Beratung zu Metadaten: Konkrete Hilfe für die Umsetzung*
  - *Konkrete Lösung für langfristige Archivierung von grossen Datenmengen nach dem FAIR Data Prinzip anbieten*
  - *technische Fragen (z.B. Metadaten, Standards, langfristige Archivierung)*
  - *Klare Information bzw. Vorgaben über die Anforderungen und Standardisierte Anforderungen bzw. klare Empfehlungen wie konkret umsetzen*
  - *Klare Definition des Begriffs Forschungsdatenmanagement*
  - *Für Publikation: Auf ein oder zwei Verlage festlegen?*

7.2 Präsentation zur Recherche über kommerzielle Big Data Cloud Plattformen

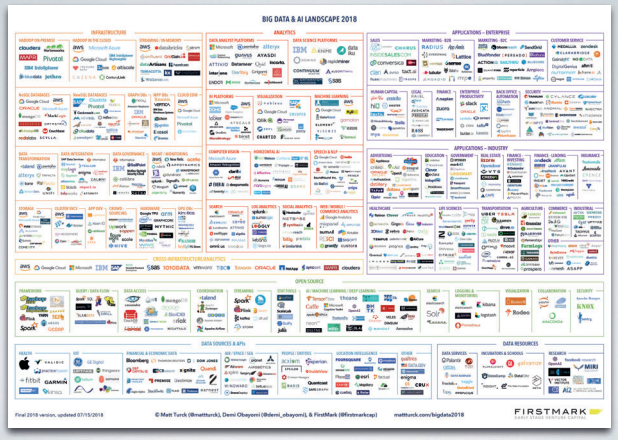
BIG DATA CLOUD PLATTFORMEN

HSLU ITC Big Data Cloud - 8.5.2019

WENN MAN NACH BIG DATA PLATTFORM GOOGELT…

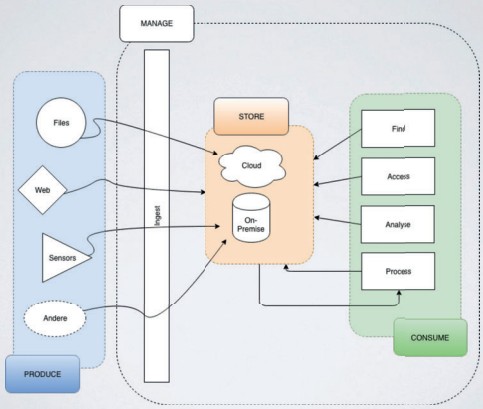


ODER NACH BIG DATA LANDSCAPE…



BACKGROUND

PROZESSÜBERSICHT



KONZEPTE

- Hadoop: Open Source Big Data Framework
- HDFS: Verteiltes Dateisystem, Teil von Hadoop
- Object Store: Speicherarchitektur, z.B Amazon S3
- Data Warehouse: Speicherung gemäss Schemata
- Data Lake: Speicherung von Rohdaten ohne Schema

PARADIGMENWECHSEL

- **Hadoop** mit **HDFS** lange Zeit dominant. Zuerst On-Premise, später auch als Cloud-Dienst.
- **Object Store** (S3) in der Cloud mit Cloud Access Layer wird immer wichtiger.

OBJECT STORE MIT CLOUD ACCESS LAYER

- "Distributed cloud-scale databases and applications on top of container infrastructure"
- Analyse Tools typischerweise einfacher zu verwenden als Hadoop Tools. SQL oder SQL-Like Syntax.
- Hadoop Tools via Compatibility Layer, falls nötig
- Beispiele: Amazon Athena, MS Cosmos, Google Big Query

NACHTEILE VON HADOOP

- Komplex und teuer (im Vergleich zu S3)
- Koppelung von Storage und Compute
- Nicht elastisch
- Stark auf Java / Scala ausgerichtet

FAZIT HDFS VS S3

- HDFS ist in der Cloud nicht (mehr) sinnvoll
- On-Premise bleibt es aber vorerst dominant, es gibt noch keine Plattformunterstützung für Private Clouds (OpenStack)

PLATTFORMEN

#1 HADOOP ON-PREMISE

- Cloudera / Hortonworks
- MapR

CLOUDERA / HORTONWORKS

- Marktführer
- Free, Open-Source Core (*CDH*)
- Simple (*Essentials*) bis umfassende (*Enterprise Data Hub*) Produktausprägungen möglich

MAPR

- Braucht kein HDFS sondern eine proprietäre Weiterentwicklung
- Ersetzt einzelne Open-Source Tools des Hadoop Stacks durch proprietäre, performance-optimierte Eigenentwicklungen

#2 HADOOP IN DER CLOUD

- Cloudera, MapR mit AWS, GCP, Azure oder anderen Cloud Providern
- Amazon Elastic MapReduce
- Azure Datalake Generation 1
- Metal Cloud Data Lake

AMAZON ELASTIC MAP REDUCE

- Amazons Version von Hadoop
- Unterstützt HDFS und S3

AZURE DATA LAKE GEN 1

- Die erste Generation von Azure Data Lake läuft mit HDFS
- Eingebunden in die Analytics- und Entwicklungswelt von Microsoft

BIGSTEP DATA LAKE

- Data Lake as a Service
- Funktioniert mit den gängigen Hadoop Tools

#3 CLOUD OBJECT STORE

- Dominiert durch die Big 3
- Speicherung, Verwaltung und Verarbeitung mit den jeweiligen Anbietertools
- Bieten aber auch Support für Hadoop Tools wie Spark

AMAZON WEB SERVICES

- Erster und grösster Anbieter
- Dutzende verschiedener Dienste
- Schwache Hybrid-Cloud Unterstützung

MICROSOFT AZURE CLOUD

- Zweitgrösster und zweiter Anbieter am Markt
- Grosses Angebot
- Gute Integration mit Microsoft Diensten

GOOGLE CLOUD PLATFORM

- Jüngster und kleinster Anbieter
- Deutlich weniger Dienste als die anderen beiden
- Kompetitive Preise
- Starker Fokus auf Machine Learning

WRAPPER

- Produkte, die als Backend AWS, Azure oder GCP verwenden…
- …und die Verwendung dieser Plattformen einfacher machen

PANOPLY

- Cloud Data Warehouse
- Bindet Data Lakes und andere Quellen komfortabel an
- Daten werden auf S3 gespeichert und mit AWS Ressourcen verwaltet
- Teuer für grosse Datenmengen und viele Quellen

WEITERE ANBIETER

- Oracle
- IBM
- SAP
- Talend
- Teradata
- Informatica

AWS PLATTFORM

KOMPONENTEN

- Laden und Katalogisierung: Glue
- Speicher: S3
- Backup: Glacier
- Zugriffskontrolle: IAM und KMS
- Wrapper: Lake Formation (beta)
- Analyse: Redshift (DWH), DynamoDB (NoSQL), Athena (SQL), EMR (Hadoop), Kinesis (Streaming), SageMaker (ML), QuickSight (BI)

LINKS

- Data Lake mit AWS
- Data Lake mit GCP
- Data Lake mit Azure
- Google Big Data Dienste
- Data Lakes and their key properties
- The case for Data Lakes
- Paradigm Shift (from Hadoop to Object Store)

DISKUSSION

ORGANISATION

- **Daten**
  - Menge / Grösse
  - Art der Quellen
  - Sensitivität
- **Infrastruktur**
  - Cloud?
  - Wer betreibt?
  - Wer verwaltet?
  - Wie wird abgerechnet?

MINIMALES FEATURESET

- Manuelle Speicherung von Rohdaten & Metadaten
- Katalogisierung
- Suche
- Download von Rohdaten

MEHR FEATURES

- Konnektoren
- Zugriffsbeschränkung
- Verschlüsselung
- Online-Analyse
- Lakeshores

NOCH MEHR FEATURES

- Self-Service
- Versionierung
- High Performance, Streaming
- Dashboards

7.3 Anforderungen für BIG DATA Anwendungen der am ITC BDC beteiligten Projekte

Projekt	**APID**	**IGE Monitoring DB**	**Low Cost Emissions-Sensor**
Projektleiter	Erny Niederberger	Reto Marek	Peter Zotter
Partner	Swisens AG	-	
Produkt	Swisens Poleno	Monitoring Datenbank für Gebäude-Energiedaten welche für wissenschaftliche Auswertungen zugänglich sein sollten.	
Gebiet	Automatische Identifikation von Pollen in Echtzeit	Gebäude/Energie-Monitoring	Emissions Messung von Holzfeuerungsabgasen
Generierte Datenmenge	Typisch 3 GByte generierte Datenmenge von einem Gerät pro Tag. Aktuell fünf Geräte im Feld. Total also 15 GByte Daten pro Tag.	Typisch 400 Datenpunkte pro Projekt, Datenpunkte werden im 1 bis 15min Intervall abgefragt. Ergibt ca. 250'000 Zeit-Wertepaare pro Tag. Über eine durchschnittliche Projektdauer von 2 Jahren ergibt dies ca. 175 Mio Zeit-Wertepaare pro Projekt. Daten werden nach Projektende allenfalls weiter geloggt. Pro Jahr rechnen wir anfangs mit 2 solcher Projekte.	Ca. 5 MB pro Tag und Messstandort
BigData Anwendung 1	Identifikation der Partikel basierend auf holografischen Bildern und Messwerten. Klassifizierung mit Hilfe von Artificial Neural Network (ANN) für (supervised und unsupervised learning)	Diverse energetische Analysen, kein wirkliches Big Data. Abfragen müssen aus R, Matlab, Tableau etc. möglich sein.	Erkennen von Optimierungspotential im Feuerungsbetrieb --> Optimierung von Regelparameter --> weniger Brennstoffeinsatz u. Emissionen
BigData Anwendung 2	Analyse der Messdaten: z.B. Holografie Rekonstruktion, Clusteranalyse	Im Zusammenspiel mit Metadaten will man automatisierte Berechnungen/Datenplausibilisierungen/Live-Checks und Analysen ermöglichen. Zur Zeit wird dies mit einer strukturierten sowie grafischen Datenbank abgebildet. Ein Algorithmus sucht sich je nach Analyse die notwendigen Datenpunkte und frägt bei der Zeitreihen-Datenbank den erforderlichen Zeitraum an um Roh- oder aggregierte Daten dieses Bereiches für die Berechnung zu kriegen.	Automatisches Erkennen von Wartungnotwendigkeit, Automatisches Erkennen hoher Emissionen und der Einhaltung der LRV
HSLU BigData Aktivitäten	17 HS P. Bächler T&A 18 FS PAIND Y. Zeder T&A 18 HS BDA Y. Zeder T&A 18 HS PAIND M. vonFlüe T&A 19 FS BDA A. Schmied T&A 19 FS BDA S. Bertuzzi I	18 Reto Marek 18 Stefan Ineichen 18 Curdin Derungs Infrastuktur wird zur Zeit in einem Projekt aktiv genutzt bis Ende 2020.	Nur projektbezogene Arbeiten von Adrian Lauber und Peter Zotter

Projekt	APID	IGE Monitoring DB	Low Cost Emissions-Sensor
Tools für die Identifikation	Machine Learning: Modelle welche mittels TensorFlow auf die Messdaten angewendet werden, Phyton	R, Matlab, Tableau, Influx Kapacitor	Phyton
Infrastruktur: Verwendete Betriebssysteme	- Linux für Datenbank und Machine-Learning - eigenes WebAPI für queries	dockerized Linux Containers für: - influxDB - Kapacitor - Telegraf - Chronograf - Grafana - neo4j - eigenes WebAPI für queries - python-Code-Executer für Imports - nodered für API-Calls - MariaDB -> Testing - Cortezo -> Testing	dockerized Linux Containers für: - influxDB - Grafana
Infrastruktur: Datenbank Provider für Messdaten	Cloudspeicher, gemietet von einem Host-Provider in der Schweiz (Nicht speziell für BigData)	Zur Zeit auf lokalem NAS, Portierung zu Switch in Planung	SWITCH Drive und SWITCH Engines
Datenbank / Object Store	SQL basiert, Zugriff über SSH und REST-API	Zeitreihen (InfluxDB) Metadaten grafisch (neo4j) Metadaten SQL-Basiert (MariaDB) -> Testphase	Influx DB (Zeitreihen optimierte DB)
Trainingsdatensatz für die Identifikation	Im Labor generierte Datensätze von verschiedenen Pollen	-	-
Umfang Trainingsdatensatz	Aktuell ca. 20 GByte wird aber laufend vergrössert	-	-
Benötigte Rechenzeit für ANN Training	> 12 Stunden für Generierung ANN auf PC mit 8-Kern CPU und relativ guter GPU	-	-
Infrastruktur: Benötigter CPU Speicher	Aktuell: 16 GByte	Aktuell ein NAS mit: Intel Celeron CPU N3160 @ 1.6 GHz 8 GB RAM -> mehr erwünscht	Aktuell: - 1 CPU, 1 VM, 1 GB RAM, 20 GB Disk, 1 IPv4. Zukunft: Evtl. mehr RAM u. CPU
Infrastruktur: Benötigter Datenspeicher und Typ	100GB SSD für Datenbank / Object Store: Für Training als lokale Files (h5)	Aktuell: 6 GB HDD für Gesamtsystem 3.8 GB davon sind Daten	noch nicht definiert, jedoch sicherlich weit unter 1TB
Wunsch Infrastruktur ANN Training	2080 TI äquivalente GPU oder besser, 64GB RAM	-	-
Wunsch Infrastruktur Fileserver	Fileserver mit Datenbank für Messdaten und Trainingsdaten 10 TByte	Fileserver mit Datenbank Zeitreihen, SQL und Graph mit 1TByte	Fileserver mit Datenbank Zeitreihen, SQL und Graph mit 1TByte

7.4 Vortrag am Forschungsplenum vom 20.08.2020 mit dem Beispiel einer Big Data Datenablage inklusive open-source-basierter Analyse- und Visualisierungstools

BigData at Swisens

12/08/2020 www.swisens.ch

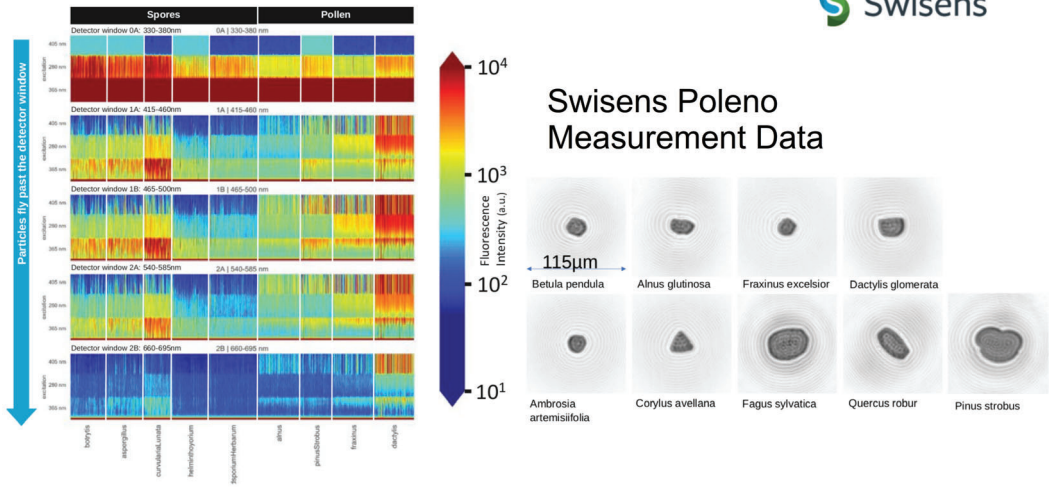
Swisens

Content

- Swisens Poleno measurement data
- Goals of data analysis
- Road to usable BigData

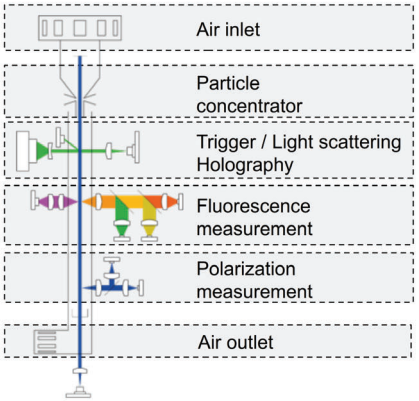
Swisens

Swisens Poleno Measurement Data



Swisens

Swisens Poleno with holography and fluorescence setup



Sampling

- Particles Ø 0.5µm to 300µm
- Flow rate: 40 l/min
- Concentration factor 1000 (particles Ø 10µm to 300µm)
- Up to 30'000 particles per m^3
- Sigma 2 sampler
- Insertable sample collector

Goals of data analysis

- Improve capabilities of device
- Gather device knowledge: Leads to better processes and future products
- Allow research usage

The Road to usable BigData

- Recording of massive data amounts *
- Storage of large amounts of data *
- Transfer form source to server infrastructure *
- Initial analysis and **filtering ***
- Comprehensive data reporting
- Make better classification models *
- Enable data sharing for community *

>>> Use available frameworks!!

The Road to usable BigData – Recording and storage

- Variable rate of data acquisition
- Data integrity guarantees (ACID)
- Performance in write AND read
- Mature and Reliable
- Data transfer integration
- Good integration in other frameworks

Running on Swisens Poleno Measurement System (Dockerized)

MySQL — SQL database

SQLAlchemy — ORM database interface

python 3 — Data processing

TensorFlow Extended — Data classification

The Road to usable BigData – Transfer

Poleno Network (VPN: OpenVPN)

- 1 TB local data buffer
- Replication delay <1s
- Fully ACID compliant
- Rest API
- Web data interface
- SSH and VNC access

MySQL Multi Master Replication

MySQL

Swisens Server

- 16 TB (currently)
- DB container
- Web data interf.
- REST API
- Offsite backups

Swisens

The Road to usable BigData – Initial Analysis and Filtering

- Fast and easy to use tool for initial look at data
- Data filtering capabilities for experts, not only developers
- Don't prohibit more advanced analysis

Data Analyzer Toolset

Swisens Data Explorer

- Fully web based
- Only need browser
- Fast first look
- Integrated filter tools
- Provides Rest API
- Handles ex- and import

Jupyter notebooks

- Full freedom
- Advanced analysis
- Custom python modules
- Handling huge data sets
- Examples are provided

Database

Swisens

The Road to usable BigData – Train Classifier Models

- Code as little as possible
- Work with very large datasets (>60GB)
- Automated data preparation
- Repeatable results
- Traceable progress

Swisens Data Explorer

- Fully web based
- Only need browser
- Fast first look
- Integrated filter tools
- Provides Rest API
- Handles ex- and import

+

ML Extension (upcoming)

- Dask handles massive data
- Tensorflow handles the training
- User chooses:
  - Known good model structure
  - Datasets to train on

Swisens

The Road to usable BigData – Data Sharing

- Allow import and export of shared files
- But, not everyone likes pre-done frameworks: No proprietary format
- Human readable format
- Future-prove – What is 10 years from now?

>> We settled on PNG, JSON and ZIP

Swisens

Thanks for your attention!
Questions?