CLOUD4EO - SCHNELLE UND EFFIZIENTE VERARBEITUNG VON ERDBEOBACHTUNGSDATEN

Neue Perspektiven der Erdbeobachtung, 2. Symposium zur angewandten Satellitenerdbeobachtung
12. bis 13. November 2019, Köln

Dr. Eva Klien

Fraunhofer-Institut für Graphische Datenverarbeitung IGD
Fraunhoferstr. 5
64283 Darmstadt

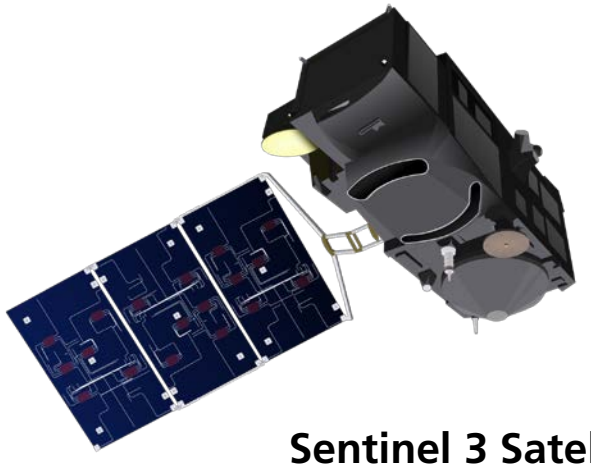
Fraunhofer IGD

M.O.S.S.
Computer Grafik Systeme
Geoinformationssysteme

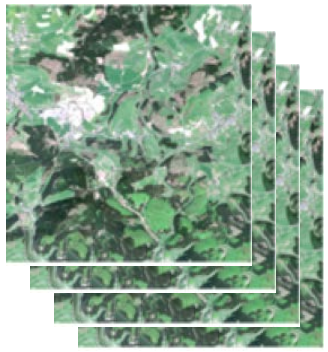
© Fraunhofer IGD

Motivation

Bild: ESA



Sentinel 3 Satellit
(Bild: SkywalkerPL, wikipedia.de)



Server-/Cloud-Infrastruktur

Bild: Paul Hammond

© Fraunhofer IGD

Fragestellung:

Wie kann eine verbesserte Nutzbarkeit und Wertsteigerung erreicht werden?

© Fraunhofer IGD

Ziele

1. Analyseverfahren zur Auswertung von Copernicus-Daten sollen effizient und skalierbar in der Cloud ausgeführt werden

2. Zugang zu den Daten und der Cloud-Infrastruktur soll erleichtert werden

© Fraunhofer IGD

Fraunhofer IGD

M.O.S.S.
Computer Grafik Systeme
Geoinformationssysteme

HESSEN

HLBG
(assoziierter Partner)

Laufzeit: November 2017-April 2019

© Fraunhofer IGD

Agenda

- Pilotanwendung
- Architekturdesign
- Umsetzung in verschiedenen Cloud-Umgebungen
- Benchmarking
- Ableitung von Handlungsempfehlungen

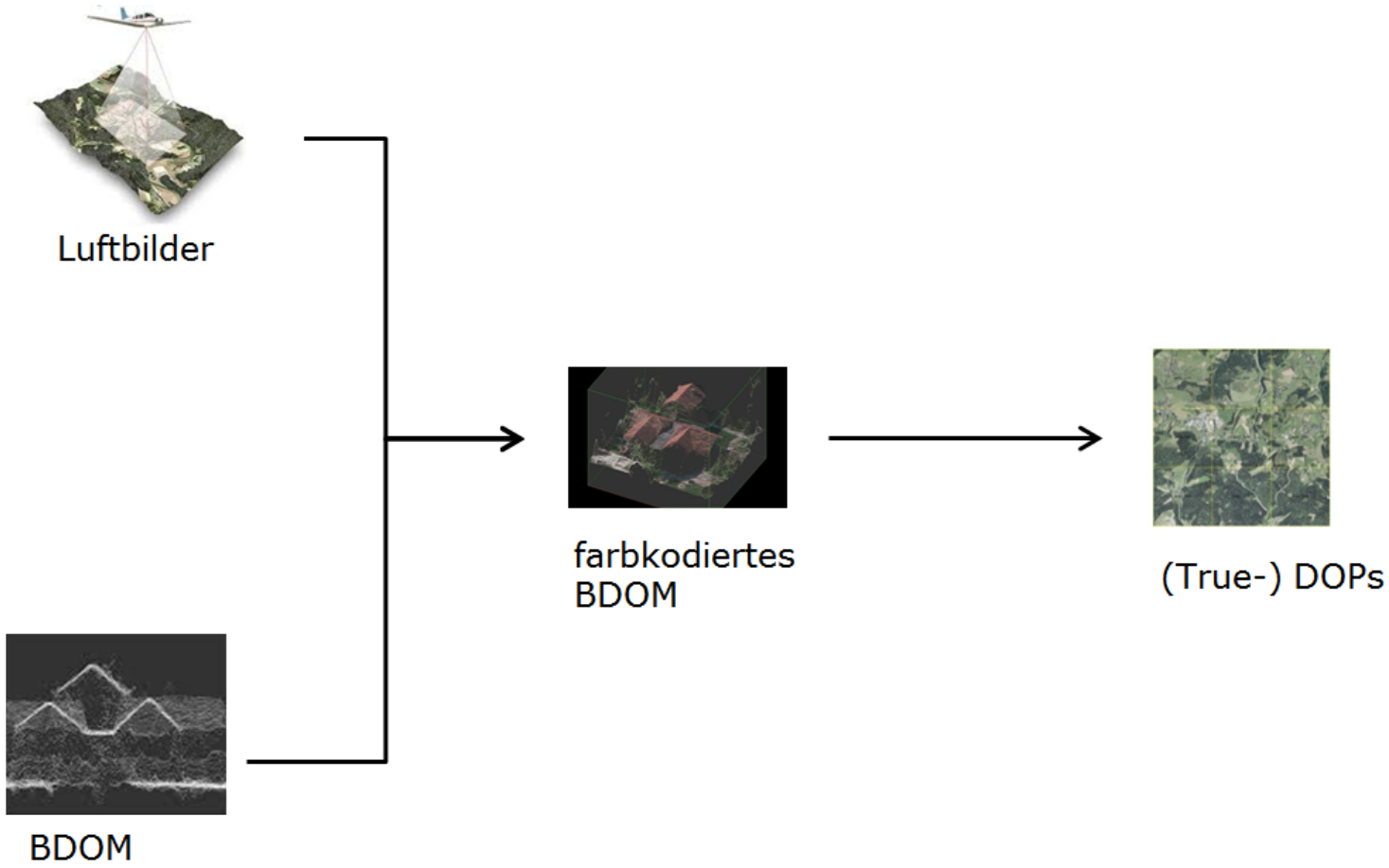
© Fraunhofer IGD

Agenda

- **Pilotanwendung**
- Architekturdesign
- Umsetzung in verschiedenen Cloud-Umgebungen
- Benchmarking
- Ableitung von Handlungsempfehlungen

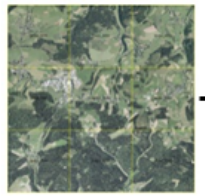
© Fraunhofer IGD

Pilotanwendung
Schritt 1: Vom bildbasierten Oberflächenmodell zum True-DOP



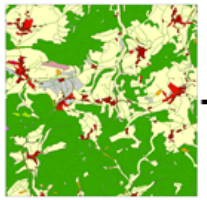
© Fraunhofer IGD

Pilotanwendung –
Schritt 2: Veränderungsdetektion

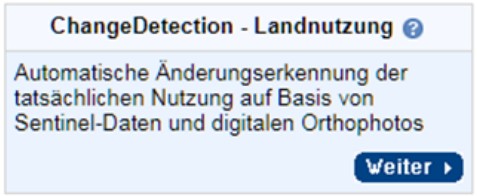


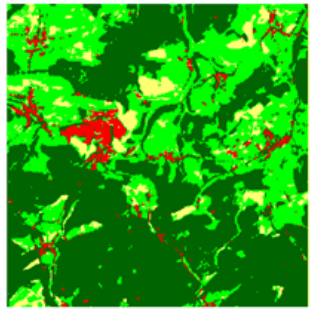
(True-) DOPs

Sentinel-2



ATKIS-DLM





Landnutzungs-klassifizierung

Fortführungs-hinweise

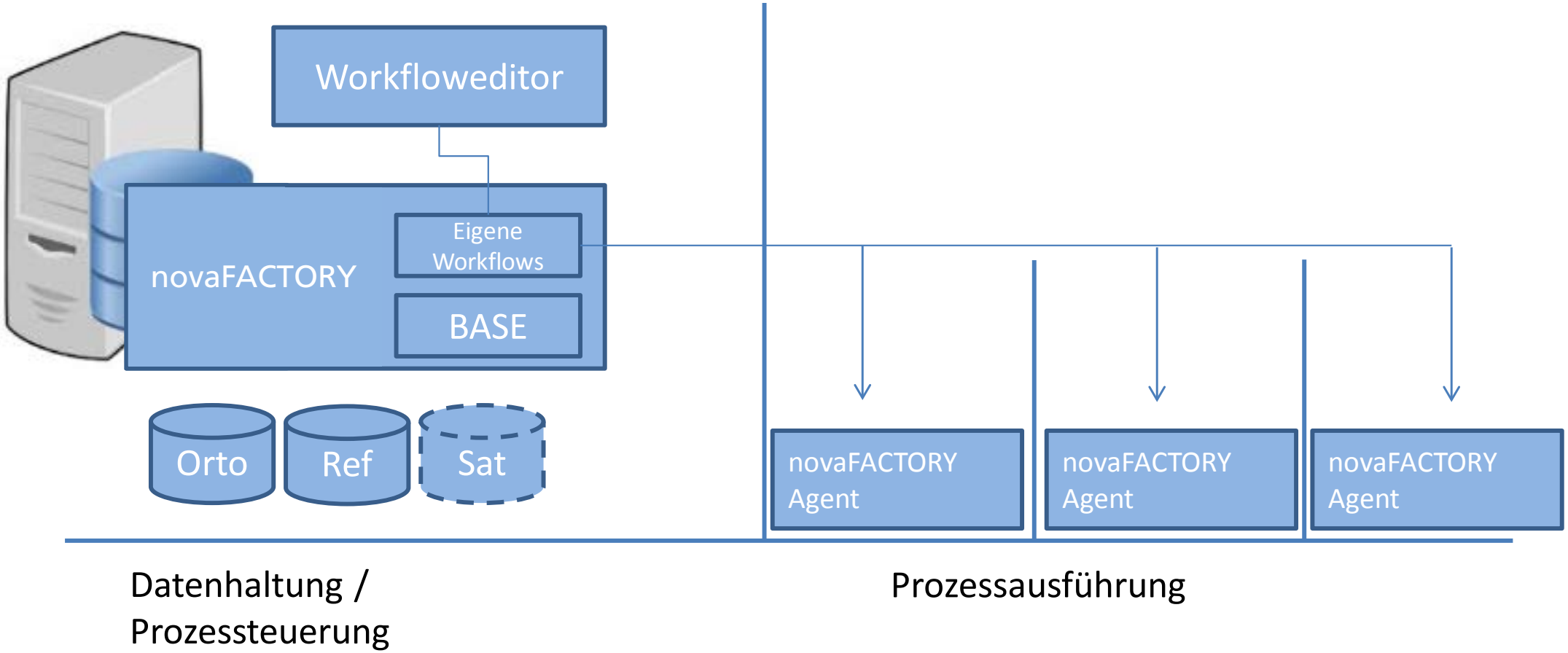
© Fraunhofer IGD

Agenda

- Pilotanwendung
- **Architekturdesign**
- Umsetzung in verschiedenen Cloud-Umgebungen
- Benchmarking
- Ableitung von Handlungsempfehlungen

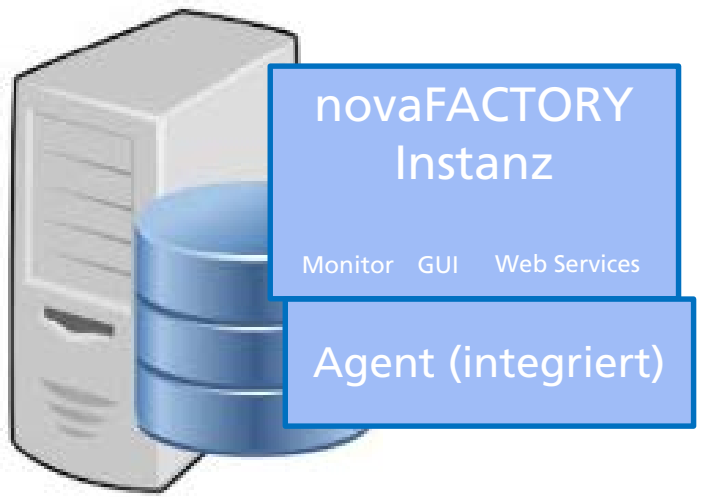
© Fraunhofer IGD

Cloud4EO-Architektur



© Fraunhofer IGD

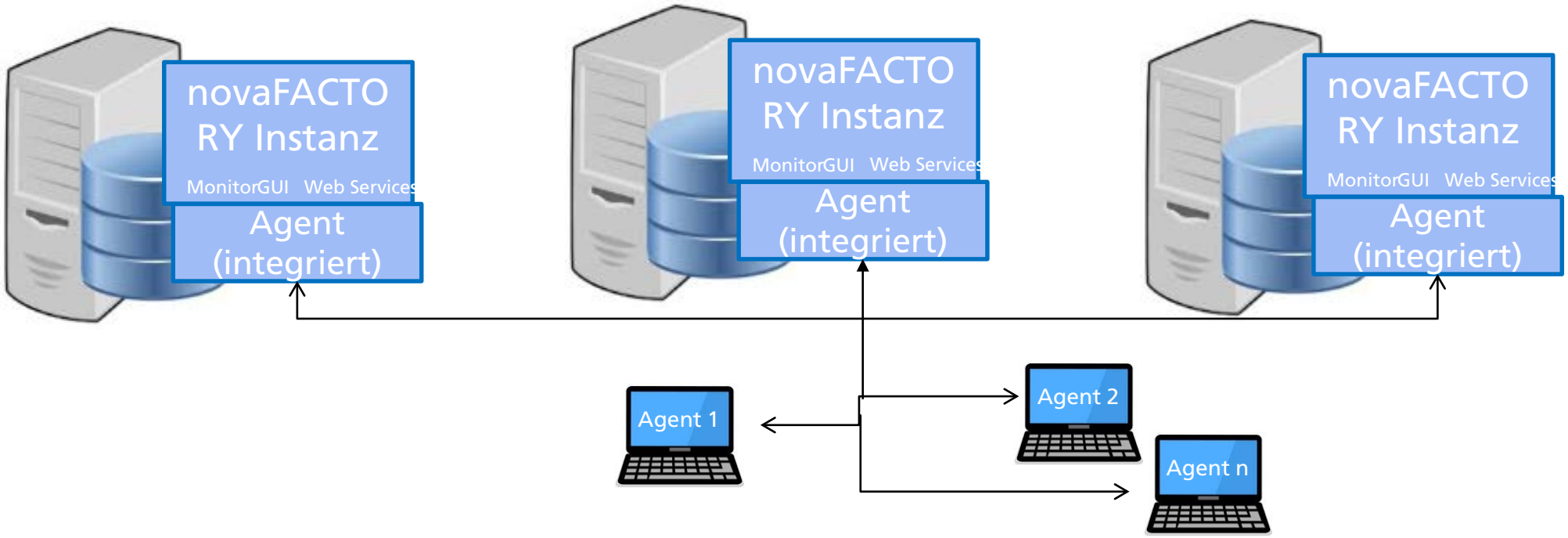
Datenhaltung / Prozessteuerung



- Unabhängige novaFACTORY Instanz (definierter Funktionsumfang) mit eigenen Workflows zur Steuerung lokal vorhandener Funktionen

© Fraunhofer IGD

Architektur



- Agent als Bestandteil von novaFACTORY unterstützt die Kommunikation zwischen den Instanzen
- Unabhängige Agent(s) ermöglichen die Anbindung von unabhängigen Ressourcen (Funktionen, Rechenleistung, Zielsystemen usw.)

© Fraunhofer IGD

Agenda

- Pilotanwendung
- Architekturdesign
- **Umsetzung in verschiedenen Cloud-Umgebungen**
- Benchmarking
- Ableitung von Handlungsempfehlungen

© Fraunhofer IGD

Anforderungen an Cloud-Umgebungen

Kriterien:

- Virtuelle Maschinen können erstellt werden
- Unternehmensstandort
- Standort des bzw. der Rechenzentren
- Datenanlieferung per Datenträger ist möglich
- Zertifizierung ist vorhanden
- Preis auf Basis einer vorab festgelegten Infrastruktur
- Windows als Betriebssystem

© Fraunhofer IGD

Anforderungen an Cloud-Umgebungen und Auswahl der Testumgebungen

Kriterien:

- Virtuelle Maschinen können erstellt werden
- Unternehmensstandort
- Standort des bzw. der Rechenzentren
- Datenanlieferung per Datenträger ist möglich
- Zertifizierung ist vorhanden
- Preis auf Basis einer vorab festgelegten Infrastruktur
- Windows als Betriebssystem

Testumgebungen:

- Reallabor (bei der M.O.S.S. GmbH)
  - Installation auf Server(n), wie sie häufig in der Praxis anzufinden ist
- Fraunhofer Cloud
  - Private OpenStack-Cloud, die vom Fraunhofer IGD betrieben wird
- Open Telekom Cloud (OTC)
  - Public Cloud mit Anbindung an CODE-DE

© Fraunhofer IGD

Rahmenwerk für Benchmarking

Für die Testszenarien wurden folgende Parameter variiert:

- Testgebiet und dadurch bedingt Datenvolumen
- Parallelisierung über mehrere Berechnungsknoten
- Parallelisierung auf einem Berechnungsknoten
- Konfiguration Berechnungsknoten

Relevant für Schritt 2 der Pilotanwendung (Änderungsdetektion):

Testszenario (Gebietsgrößen)	Datenvolumen ATKIS (SHP-Format)	Datenvolumen DOP (RGBI, 20cm)	Datenvolumen Sentinel-2
V_1	160 KB	400 MB	250 GB
V_2	4 MB	10 GB	250 GB
V_3	40 MB	100 GB	250 GB
V_4	120 MB	400 GB	250 GB
V_5	500 MB	2 TB	250 GB

Testszenario (Parallelisierung)	Anzahl paralleler Prozesse
P1_1	1
P1_2	2
P1_4	4
P3_1	3
P3_2	6
P3_4	12

© Fraunhofer IGD

Fraunhofer IGD

Agenda

- Pilotanwendung
- Architekturdesign
- Umsetzung in verschiedenen Cloud-Umgebungen
- **Benchmarking**
- Ableitung von Handlungsempfehlungen

© Fraunhofer IGD

Benchmarking Ergebnisse

Ausführliche Benchmarktests

auf Basis der festgelegten Testszenarien

in den drei unterschiedlichen Testumgebungen:

Volumen/ Gebiet	Parallelisierung	Ausführungs-umgebung	Gesamtlaufzeit [hh:mm:ss.ms]	Paralleleffekt in %
V_1	P1_1	Fraunhofer Cloud	00:16:23	99,96%
V_2	P1_1	Fraunhofer Cloud	01:24:49	99,98%
V_2	P1_2	Fraunhofer Cloud	00:59:59	189,87%
V_2	P1_4	Fraunhofer Cloud	00:55:18	374,44%
V_2	P3_1	Fraunhofer Cloud	00:29:00	274,90%
V_2	P3_2	Fraunhofer Cloud	00:22:49	469,55%
V_3	P1_1	Fraunhofer Cloud	07:30:54	99,97%
V_3	P1_2	Fraunhofer Cloud	10:22:45	197,40%
V_3	P1_4	Fraunhofer Cloud	04:42:16	384,21%
V_3	P3_1	Fraunhofer Cloud	02:29:18	297,20%
V_3	P3_2	Fraunhofer Cloud	01:52:10	565,19%
V_3	P3_4	Fraunhofer Cloud	01:47:48	983,63%
V_4	P1_1	Fraunhofer Cloud	18:49:20	99,96%
V_4	P1_2	Fraunhofer Cloud	12:45:54	199,54%
V_4	P1_4	Fraunhofer Cloud	21:28:30	395,89%
V_4	P3_1	Fraunhofer Cloud	06:06:52	296,68%
V_4	P3_2	Fraunhofer Cloud	04:27:37	570,99%
V_4	P3_4	Fraunhofer Cloud	03:54:06	1112,41%
V_1	P1_1	Reallabor	00:09:56	99,92%
V_2	P1_1	Reallabor	02:08:10	99,98%
V_2	P1_2	Reallabor	01:52:50	196,83%
V_2	P1_4	Reallabor	00:58:35	250,50%
V_2	P3_1	Reallabor	00:47:23	311,31%
V_2	P3_2	Reallabor	00:50:13	264,49%
V_3	P1_1	Reallabor	11:26:43	99,98%
V_3	P1_2	Reallabor	07:56:18	197,38%
V_3	P1_4	Reallabor	06:39:40	383,35%
V_3	P3_1	Reallabor	11:29:48	99,93%
V_3	P3_2	Reallabor	04:33:48	290,67%
V_4	P1_1	Reallabor	39:38:52	99,95%
V_4	P1_2	Reallabor	33:45:44	199,74%
V_4	P1_4	Reallabor	15:57:25	387,67%
V_4	P4_1	Reallabor	07:15:25	378,34%
V_4	P3_4	Reallabor	04:58:49	714,15%
V_1	P1_1	OTC	18:09:47	100,00%
V_2	P1_1	OTC	21:26:34	205,86%
V_2	P1_2	OTC	20:29:23	246,10%
V_2	P1_4	OTC	21:23:59	189,16%
V_2	P3_1	OTC	01:31:44	220,86%
V_2	P3_2	OTC	01:33:02	339.52%
V_3	P1_4	OTC	03:57:04	247,14%
V_3	P3_1	OTC	03:26:50	131,72%
V_3	P3_2	OTC	14:00:52	178,19%
V_3	P3_4	OTC	13:37:41	278,33%
V_4	P3_1	OTC	38:47:42	188,71%
V_4	P3_2	OTC	29:06:10	322,86%
V_4	P3_4	OTC	23:10:26	597,92%

© Fraunhofer IGD

Benchmarking Ergebnisse

Volumen/ Gebiet	Parallelisierung	Ausführungs-umgebung	Gesamtlaufzeit [hh:mm:ss.ms]	Paralleleffekt in %
V_1	P1_1	Fraunhofer Cloud	00:16:23	99,96%
V_2	P1_1	Fraunhofer Cloud	01:24:49	99,98%
V_2	P1_4	Fraunhofer Cloud	00:55:18	379,44%
V_2	P3_2	Fraunhofer Cloud	00:22:49	469,55%
V_3	P1_2	Fraunhofer Cloud	10:22:45	197,40%
V_3	P1_4	Fraunhofer Cloud	04:42:16	384,21%
V_3	P3_1	Fraunhofer Cloud	02:29:18	297,20%
V_4	P1_4	Fraunhofer Cloud	21:28:30	395,89%
V_4	P3_2	Fraunhofer Cloud	04:27:37	570,99%
V_4	P3_4	Fraunhofer Cloud	03:54:06	1112,41%
V_1	P1_1	Reallabor	00:09:56	99,92%
V_2	P1_1	Reallabor	02:08:10	99,98%
V_2	P1_2	Reallabor	01:52:50	196,83%
V_2	P1_4	Reallabor	00:58:35	390,50%
V_2	P3_1	Reallabor	00:47:23	469,55%
V_2	P3_2	Reallabor	00:50:13	264,49%
V_3	P1_1	Reallabor	11:26:43	99,98%
V_3	P1_2	Reallabor	07:56:18	197,38%
V_3	P1_4	Reallabor	06:39:40	383,35%
V_3	P3_1	Reallabor	11:29:48	99,93%
V_3	P3_2	Reallabor	07:30:54	290,67%
V_4	P1_1	Reallabor	39:38:52	99,95%
V_4	P1_2	Reallabor	33:45:44	199,74%
V_4	P1_4	Reallabor	15:57:25	387,67%
V_4	P4_1	Reallabor	07:15:25	378,34%
V_4	P3_4	Reallabor	04:58:49	714,15%
V_1	P1_1	OTC	18:09:47	100,00%
V_2	P1_1	OTC	21:26:34	205,86%
V_2	P1_2	OTC	20:29:23	246,10%
V_2	P1_4	OTC	21:23:59	189,16%
V_2	P3_1	OTC	01:31:44	220,86%
V_2	P3_2	OTC	01:33:02	339,52%
V_3	P1_4	OTC	03:57:04	247,14%
V_3	P3_1	OTC	03:26:50	131,72%
V_3	P3_2	OTC	14:00:52	178,19%
V_3	P3_4	OTC	13:37:41	278,33%
V_4	P3_1	OTC	38:47:42	188,71%

→ **Parallelisierung von Prozessen bei einer geeigneten Aufteilung bzw. Kachelung der Daten bringt Vorteile in der Ausführungszeit in allen Ausführungsumgebungen**

Nachteile vor allem aufgrund von Windows als Betriebssystem:

→ standardisierte Cloudtechnologien konnte vor allem in der Softwareverteilung (wie Ansible, Chef oder Puppet) nicht bzw. nicht optimal genutzt werden können

→ komplexere Verteilung von Betriebssystemimages auf mehrere Knoten gegenüber anderen Betriebssystemen wie z.B. Linux

Detaillierte Informationen zu Testumgebungen, Testkonfiguration, Ergebnissen und vergleichender Bewertung finden sich im veröffentlichten Abschlussbericht

© Fraunhofer IGD

Nutzen von Copernicus-Daten für das Anwendungssegment

Hoher Nutzen durch:

- hohe Aktualität der möglichen Veränderungsdetektion
- hohe zeitliche Variabilität ermöglicht die Unterscheidung von Objektklassen z.B. über den dezidierten Vergleich von Vegetationsständen in Frühjahrs-, Sommer und Herbstbildern → Grundlage für eine bessere Landnutzungsklassifizierung
- hohe radiometrische Auflösung ermöglich ebenfalls eine bessere Unterscheidung und Klassifizierung von Objektklassen (über die in DOPs üblicherweise vorliegenden vier Kanäle RGBI hinaus)

Nachteile:

- durch relativ niedrige geometrische Auflösung der Sentinel-2 Daten (10x10m) können viele Objekte der Landnutzung nicht ausreichend sicher klassifiziert werden. Daher ist die parallele Nutzung der amtlichen DOPs mit einer Bodenauflösung von 20x20cm für den Gesamtprozess unerlässlich
- potentielle Verbesserung der Gesamtergebnisse durch die Hinzunahme von Sentinel-1 Radardaten möglich

© Fraunhofer IGD

Agenda

- Pilotanwendung
- Architekturdesign
- Umsetzung in verschiedenen Cloud-Umgebungen
- Benchmarking
- **Ableitung von Handlungsempfehlungen**

© Fraunhofer IGD

Handlungsempfehlungen

- CODE-DE ermöglicht Zugriff auf Sentinel-Daten, verteilte Ausführungsumgebung, fertige Prozessierungsdienste
- Die Verbindung von der CODE-DE-Plattform mit der OTC ermöglicht eine nahezu beliebige Erhöhung der Verarbeitungskapazität
- Die Benchmarktests in diesem Projekt haben gezeigt, dass Anwendungen aus dem Erdbeobachtungsbereich von Parallelisierung und Verteilung auf verschiedene Recheneinheiten profitieren, und dass trotz Overhead und nicht zu 100% linearer Skalierung ein großer Geschwindigkeitsvorteil erreicht werden kann.

© Fraunhofer IGD

Handlungsempfehlungen…

… für den Ausbau einer Verarbeitungsplattform wie CODE-DE:

→ Ausführungsumgebung von CODE-DE hinsichtlich Windows und der Unterstützung kommerzieller Software erweitern

Hintergrund:

- Viele kommerzielle Produkte zu Verarbeitung mit Erdbeobachtungsdaten sind derzeit nur mit Windows kompatibel

© Fraunhofer IGD

Handlungsempfehlungen…

… für den Ausbau einer Verarbeitungsplattform wie CODE-DE:

→ die Möglichkeit anbieten, dass Anwender Daten per Festplatte oder USB-Stick auf dem Postweg über einen definierten Prozess versenden können, die dann vor Ort in die Verarbeitungsplattform importiert werden

Hintergrund:

- Für viele Anwendungen werden zusätzlich zu dem direkten Zugriff auf die Sentinel-Daten noch weitere, ebenfalls sehr große Daten erforderlich
- Die Anbindung aus dem DFN an die OTC ist sehr gut, kann aber nicht von allen Anwendern und für alle Daten genutzt werden
- In der Amazon Cloud AWS steht mit dem Produkt „Snowball“ bereits ein ähnliches Konzept zur Verfügung (https://aws.amazon.com/de/snowball/)

© Fraunhofer IGD

Handlungsempfehlungen…

… für den Ausbau einer Verarbeitungsplattform wie CODE-DE:

→ ein Management System für Scientific Workflows in CODE-DE integrieren, um automatisierte komplexe, verteilte Arbeitsabläufe zu ermöglichen

Hintergrund:

- Datengetriebenen Workflows (sogenannte „Scientific Workflows“ –Verarbeitung durch mehrere, hintereinander geschaltete Prozessierungsdienste) sind sehr üblich
- CODE-DE-Plattform ermöglicht lediglich Anwendung eines einzelnen Prozessor auf die Quelldaten (Workflows müssen manuell verwaltet werden)
- Spezialisierte Workflow Management Systeme:
  - übernehmen Ausführung der Workflows vollständig automatisiert
  - können die einzelnen Prozessierungsdienste in verteilten und sich dynamisch ändernden Umgebungen auszuführen.
  - ermöglichen die Realisierung komplexer Abläufe, die elastisch mit den zur Verfügung stehenden Ressourcen skalieren und beliebig große Datenmengen verarbeiten können.

© Fraunhofer IGD

Handlungsempfehlungen…

… für den Ausbau einer Verarbeitungsplattform wie CODE-DE:

→ eine bessere Integration von CODE-DE mit OTC realisieren

Hintergrund:

- Flexible Anpassung der Verarbeitungs- und Speicher-Ressourcen an die Anforderungen der Markteilnehmer wir in Zukunft immer wichtiger werden
- Die Open Telekom Cloud (OTC) basiert auf OpenStack und ermöglicht durch etablierte Schnittstellen gute Anbindung an vorhandene Software

© Fraunhofer IGD

VIELEN DANK FÜR IHRE AUFMERKSAMKEIT!

Der Abschlussbericht des Projekts „Cloud4EO“ ist über die Technische Informationsbibliothek (TIB) erhältlich sowie über die beiden Projektpartner Fraunhofer IGD und M.O.S.S. GmbH

Dr. Eva Klien

Fraunhofer IGD
Fraunhoferstraße 5
64283 Darmstadt

Tel.: +49 (0) 6151 155 – 412
Fax.: +49 (0) 6151 155 – 444
E-Mail: eva.klien@igd.fraunhofer.de

http://www.igd.fraunhofer.de

Daniel Holweg

M.O.S.S. Computer Grafik Systeme GmbH
Hohenbrunner Weg 13
82024 Taufkirchen

Telefon: +49 89 66675-100
Telefax: +49 89 66675-180
E-Mail: info@moss.de

https://www.moss.de

© Fraunhofer IGD

Fraunhofer IGD