Data Lifecycle Lab Erde und Umwelt

LSDMA All-Hands Meeting 13. März 2013

Jörg Meyer
Parinaz Ameri, Carsten Ehbrecht, Silke Halstenberg, Stephan Kindermann, Michael Lautenschlager
DKRZ, KIT

LSDMA

Meetings und Workshops

- Nov 2012: KIT-interner Workshop DLCL „Erde und Umwelt“
  - Teilnehmer vom
    - Institut für Meteorologie und Klimaforschung - Atmosphärische Spurengase und Fernerkundung (IMK-ASF)
    - Institut für Meteorologie und Klimaforschung - Forschungsbereich Troposphäre (IMK-TRO)
    - Institut für Photogrammetrie und Fernerkundung (IPF)
    - Steinbuch Centre for Computing (SCC)
  - Ziele: neue Kontakte, detaillierteres Verständnis der DLCs

- Feb 2013: DKRZ-KIT Workshop DLCL „E&U“ und DSIT
  - Teilnehmer: DKRZ, KIT (SCC, IPE)
  - Ziele: Arbeitsplan für gemeinsame Projekte, Koordination von Aktivitäten im DLCL mit DSIT und LSDF

- Feb 2013: KIT-Treffen mit SimLab „Klima und Umwelt“

IMK-ASF - Interferometerentwicklung für Flugzeug und Ballon (IFB)

- Interferometerentwicklung für Flugzeug: GLORIA-Projekt
- Datenlebenszyklus:
  - Datennahme: Flug (ca. 10h) → 2TB Daten
  - Prozessierung in mehreren Schritten (Levels)
  - Daten werden in LSDF gespeichert
  - Zugriff auch aus Jülich erwünscht
- Projekt:
  - automatische Speicherung der Metadaten der Prozessierungskette in einer MongoDB (NoSQL)
  - mehrere Instanzen der DB (IMK, LSDF, Jülich)
  - Sekundärdaten ebenfalls in MongoDB
- Test von Datenzugriff über „domain specific data integration“ (SCC-Projekt)

- Status:
  - ~25 TB Primärdaten, ~50 TB Sekundärdaten
  - Test-MongoDB-Instanzen in LSDF-Cloud, am IMK und am Institut für Angewandte Informatik (KIT/IAI)

IMK-ASF - Fernerkundung von Spurenstoffen mit Satelliten (SAT)

- Satellitendaten (z.B. MIPAS-Gerät in Envisat)
- Datenformate und Datenlebenszyklus:
  - Datenformat von Vergleichsdaten (MLS): hdf5
  - Zwischenschritte: ASCII
  - MySQL Metadaten-DB (MIPAS):
    - Anfragen (zu) langsam
    - neue Struktur erwünscht
  - weitere Formate: netCDF, IDL (Visualisierung teilweise auch mit IDL)
- Projekt:
  - Migration der bestehenden Meta-Daten in neue DB (MongoDB)
  - DB erweitern (kein ASCII mehr in Workflow)
  - DB-Zugriffe mit MapReduce (parallel)
- Status:
  - Analyse des bestehenden SQL-Schemas und der Analyse-Tools
  - Analyse typischer Queries und Analyseschritte
  - erste DB-Zugriffsoptimierungen
- Zwischenziel:
  - Vergleich MySQL und MongoDB auf gleicher Hardware

IMK-ASF - Meteorologische Satelliten - Anwendungen (MSA)

- Satellitendaten
- Projekt:
  - Ablage von Satelliten-Daten (METEOSAT) und Wettermodellen (Quelle ECMWF) auf LSDF strukturieren
  - Metadaten-Katalog erstellen
  - (Geo-Informationssystem (GIS) als Schnittstelle)
- bisher:
  - Daten in Verzeichnisstruktur mit Namenskonvention
  - Auswertung langer Zeitreihen schwierig
  - keine gruppenübergreifende Synergien

- Status:
  - Schriftliche Beschreibung der bisherigen Datenorganisation
- Zwischenziel:
  - Erstellung eines Planes mit genauen Anforderungen und technischen Umsetzungen

Deutsches Klimarechenzentrum

- Feb 2013: KIT-DKRZ Workshop https://bscw.zam.kfa-juelich.de/bscw/nj_bscw.cgi/d898604/DKRZ-KIT-Protokoll-final.doc
- gemeinsames Projekt:
  - iRODS-Föderation KIT-DKRZ aufsetzen,
    - Replikation von Daten zwischen DKRZ und KIT (LSDF)
    - wechselseitiger Zugriff über iRODS
  - Use Case:
    - Nutzer vom IMK-TRO rechnet am DKRZ und kopiert dafür Antriebsdaten (portionsweise) ans DKRZ.
    - z.Z. manuelles Kopieren (scp)
    - Strukturierung von Workflows durch Replikation über iRODS
    - Systematische Performance/Bandbreitenmessungen zwischen iRODS-Instanzen möglich
  - Use Case II:
    - Teil 2 des Use Case: Zugriff auf die Rechenergebnisse am DKRZ
    - Ergebnisdaten am DKRZ transparent bearbeitet und reduziert
    - nur Ergebnisse werden zur Endverarbeitung zum KIT transferiert
- Status und Zwischenziel:
  - sowohl KIT als auch DKRZ haben bereits iRODS-Föderationen aufgesetzt
  - Planung und technische Vorbereitungen an beiden Standorten im Gange

Zusammenfassung

- GLORIA:
  - konkrete technische Planung und erste Tests
  - Strukturierte Zusammenarbeit
  - geplanter Test von DSDI
- Satellitendaten
  - Ziele:
    - Performance-Verbesserung bei Datenbankzugriffen (Ansatz: MySQL → MongoDB + MapReduce)
    - Workflow Optimierung
  - Status: Analyse von MySQL DB und Tools
- Metadaten-Katalog:
  - Bedarf an Datenmanagement (Kataloge, einfache Schnittstellen, …)
  - Soll Klima-Analysen vereinfachen (lange Zeitreihen)
  - Ziele und Anforderungen weiter herausarbeiten
- DKRZ:
  - Datenreplikation zwischen KIT-DKRZ (iRODS)
  - Use-Case: Anwender von IMK-TRO
  - Zusammenarbeit mit DSIT für iRODS
- SimLab: Zusammenarbeit mit IMK im HPC-Bereich;
  - Gemeinsames DLCL-HPC-Projekt angestrebt