

Geballte sächsische Big-Data-Expertise in Wissenschaft, Industrie und Wirtschaft

ScaDS Dresden/Leipzig – Ein nationales Kompetenzzentrum für die Big-Data-Forschung

ScaDS Dresden/Leipzig wird als eines von zwei nationalen Kompetenzzentren für den intelligenten Umgang mit Big Data durch das BMBF gefördert. Durch die enge Kooperation der beiden Wissenschaftsstandorte sowie zahlreiche Industrie- und Wirtschaftspartner bündelt es eindrucksvoll die sächsische Expertise auf diesem Gebiet. Damit können – von ganz konkreten Fragestellungen ausgehend – Lösungen und Dienste für die unterschiedlichen Herausforderungen im Umgang mit komplexen und immer weiter wachsenden Datenbeständen in Forschung, Industrie und Wirtschaft entwickelt werden.



Das Spektrum des Kompetenzzentrums umfasst bislang wissenschaftliche und kommerzielle An-

wendungsgebiete in den Lebens-, Material- und Ingenieurwissenschaften, Umwelt- und Verkehrswissenschaften, Digital Humanities sowie aus dem Business-Umfeld, in denen zur Lösung komplexer Fragestellungen schwerpunktmäßig Informatikmethoden aus den Bereichen Datenqualität und -integration, Wissensextraktion und visuelle Analyse eingesetzt werden. Durch intelligentes Data Life Cycle Management und angepasste Rechnerarchitekturen können großskalige Anwendungsszenarien zur Bearbeitung sehr großer und komplexer Datenmengen umgesetzt werden. Das ScaDS-Servicezentrum unterstützt die Analyse der wissenschaftlichen Fragestellungen in den unterschiedlichen Anwendungsgebieten und macht die entwickelten Lösungen darüber hinaus für andere Fachgebiete nutzbar, sodass ein möglichst breites Spektrum von Anwendern davon profitieren kann.

Im Bereich Datenintegration werden z. B. Methoden zur automatisierten Ausführung von ETL-Workflows (Extract/Transform/Load) auf großen Datenbeständen unter gleichzeitiger Duplikat-Erkennung untersucht. Weitere Schwerpunkte sind die Einhaltung sicherheitsrelevanter Anforderungen (privacy-preserving) auf anonymisierten Datenbeständen sowie die graph-basierte Datenintegration für Analysen auf großen Graphdatensätzen. Hierbei werden u. a. Methoden aus Open-Source-Analytics-Umgebungen (z. B. Apache Spark, Apache Flink) angewandt, um In-Memory-Technologien für Nahezu-Echtzeitanalysen einzusetzen.

Im Bereich Wissensextraktion stehen die Strukturierung heterogener unstrukturierter Datentypen sowie semantische Analysen im Vordergrund. Unterschiedliche Analysemethoden werden u. a. auf Text-, Bild- und Videodaten sowie Hochdurchsatz-Sequenzierungsdaten angewendet. Ein weiterer Schwerpunkt ist die Untersuchung von Deep-Learning-Ansätzen zur Annotation von Bildinhalten. Hier werden beispielsweise Algorithmen zur Siedlungserkennung und -entwicklung angewendet, um Strukturen in geografischem Kartenmaterial

semantisch zu annotieren oder Bilder aus hochauflösenden Mikroskopen zu analysieren, um durch Strukturerkennung die Entwicklung biologischer Modellorganismen zu studieren. Daneben sind die Entwicklung eines Frameworks zur skalierbaren Graphenanalyse sowie die Bereitstellung standardisierter Dienste zur Sprach- und Textanalyse aktuelle Forschungsaktivitäten.

Die visuelle Analyse unterstützt die Darstellung großer Datensätze zur Auswertung. Hierbei werden generische Methoden bereitgestellt, wie die Darstellung von punktförmigen (Teilchen)-Daten. Anwendungsgebiete sind u. a. die Material- und Ingenieurwissenschaften, in denen die Visualisierung von Simulationsdaten auf unterschiedlichen Größenskalen von großem Interesse ist oder die Darstellung von Genomdaten in den Lebenswissenschaften.

Die inhaltliche und algorithmische Forschung wird durch Methoden zur Ausführung komplexer Analyseverfahren auf großen Rechnersystemen unterstützt. Dazu stehen ScaDS Dresden/Leipzig moderne Hochleistungsrechner am ZIH der TU Dresden zur Verfügung. Dort werden umfangreiche Softwarewerkzeuge wie die generischen Analyse-Frameworks Hadoop, Spark und Flink oder Dienste zur Modellierung und Ausführung komplexer datengetriebener Analyse-Workflows eingesetzt. So können sich Nutzer – ohne direkte Interaktion mit den Hochleistungsrechnern – auf die inhaltliche Fragestellung konzentrieren, da die einzelnen Teilanalysen automatisch erzeugt und auf den Rechenressourcen verteilt werden. Tausende individuelle Tasks können so automatisch verwaltet werden. Die Ergebnisse bleiben in der Workflow-Umgebung zur weiterführenden Nutzung verfügbar. ■

Kontakt

Technische Universität Dresden
Zentrum für Informationsdienste
und Hochleistungsrechnen (ZIH)

Dr. René Jäkel
01062 Dresden

Tel.: +49 351 463-42331

rene.jaekel@tu-dresden.de

http://tu-dresden.de/zih

http://scads.de



Multi-Skalen-Visualisierung von Bauteilsimulationen
Bild: ScaDS Dresden/Leipzig