Unterstützung datenintensiver Forschung am KIT - Aktivitäten, Dienste und Erfahrungen

Bernhard Neumair, Achim Streit

Karlsruhe Institute of Technology (KIT)
Steinbuch Centre for Computing
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
bernhard.neumair@kit.edu
achim.streit@kit.edu

Datenintensive Forschung oder auch *Big Data in Science* adressiert nicht nur die Herausforderungen, die durch die schiere Größe und Menge an produzierten Daten (Volume) entstehen, sondern auch die Fragestellungen rund um die Geschwindigkeit der Entstehung der Daten (Velocity), der Vielfalt der Daten (Variety), der Wahrhaftigkeit und Authentizität der Daten (Veracity) und – nicht zu vergessen – den Wert der Daten (Value). Dabei stellt die *Exploration der Daten* eine Revolution dar, wie wissenschaftliche Erkenntnisse gewonnen werden können; dies bildet die *4. Säule der modernen Wissenschaft* neben Theorie, Experiment und Simulation. Die hohe Bedeutung der Exploration großer Datenmengen zeigt das Beispiel des Large Hadron Colliders (LHC) am CERN. Neben dem Beschleuniger und den Experimenten bildet das groß-skalige, verteilte Management von Daten und deren Analyse die Grundlage für die Entdeckung des Higgs-Teilchens.

Wissenschaftliche Fragestellungen rund um die Unterstützung datenintensiver Forschung am KIT sind: Wie kann die langfristige Erhaltung von Daten und die Bestandspflege mit Werkzeugen und Prozessen unterstützt werden? Wie können wertvolle und personenbezogene Daten geschützt werden? Wie können Daten auf einfache Art und Weise in globalen Kooperationen gemeinsam benutzt werden? Wie kann der Umgang mit Meta-Daten optimiert werden? Wie können Daten archiviert und für lange Zeit aufbewahrt werden? Wie können Erkenntnisse aus Daten mit neuartigen Methoden gewonnen werden? Und wie müssen Curricula zur Ausbildung zukünftiger Datenwissenschaftler, Dateningenieure und Datenanalysten aussehen?

Zur Adressierung der o.g. Fragestellungen und zur optimalen Unterstützung datenintensiver Forschung sind Kooperationen ein elementarer Faktor. Am KIT arbeiten wir sehr eng in gemeinsamer Forschung mit zahlreichen *Wissenschaftscommunities* zusammen und kooperieren auf *nationaler und internationaler Ebene* mit anderen Wissenschaftseinrichtungen und der Industrie.

Im Vortrag wird auf diese Beispiele, die gemeinsamen Aktivitäten, die darin entwickelten Dienste und die gewonnen Erfahrungen ausführlich eingegangen. Unter anderem werden vorgestellt: *GridKa als deutsches Tier-1 Daten- und Rechenzentrum*, das alle 4

Experimente des LHC-Beschleunigers am CERN unterstützt; die *Large Scale Data Facility (LSDF)* zur Bereitstellung von Speicher- und Analysekapazitäten für verschiedenste Forschungsrichtungen wie z.B. der Systembiologie, der Klimaforschung, der Forschung mit Synchrotronstrahlung, den Geisteswissenschaften und der Erdsystemforschung; die Entwicklung von IT-Diensten für Nutzer im Land Baden-Württemberg, z.B. zum einfachen Austausch von Daten a la Dropbox.

Darüber hinaus werden ausgewählte wissenschaftliche Highlights aus der gemeinsamen Forschung im Rahmen der *Helmholtz-Portfolioerweiterung LSDMA* anhand der modernen Lichtscheiben-Mikroskopie und der technologischen Analyse von Stromnetz-Messdaten dargestellt und auf Erfahrungen in der datenintensiven Forschung zusammen mit verschiedenen Wissenschaftscommunities eingegangen.

Ein weiterer Schwerpunkt des Vortrags ist das *Smart Data Innovation Lab (SDIL)*, das Anfang 2014 am KIT öffentlich vorgestellt wurde. Die Vision des SDIL ist der Aufbau eines bundesweiten Verbunds für *In-Memory Big Data Analyse* komplexer Datenbestände mit intelligenten Verfahren (aka Smart Data), der für die kooperative Forschung und Entwicklung durch Wirtschaft und Wissenschaft genutzt wird. Am KIT soll eine *leistungsstarke IT-Infrastruktur* von der Hardware- und Software-Industrie bereitgestellt werden, mit der die Wissenschaft gemeinsam mit der Wirtschaft hoch-performante Arbeit mit Big Data durchführen kann. Um eine realitätsnahe Forschung zu ermöglichen, liefern *Industriepartner Datenquellen aus der Praxis*, mit denen Forschung in strategisch wichtigen Feldern unterstützt wird. Diese Datenquellen werden ergänzt durch Daten der öffentlichen Hand sowie im Internet frei verfügbare Datenquellen.

Das SDIL richtet sich mit seinem Angebot zunächst an Forschungsarbeiten in den *strategisch wichtigen Feldern* Industrie 4.0, Energie, Smart Cities und Medizin. Andere Forschungsfelder werden später folgen. Um in diesen Bereichen mit einem umfassenden Angebot arbeiten zu können, wird SDIL mit *Data Innovation Communities* in jedem der strategischen Forschungsfelder arbeiten. Diese Communities werden von je einem Partner aus Wissenschaft und Wirtschaft als Moderator geleitet. Ziele der Data Innovation Communities sind die Diskussion und Definition von Forschungsschwerpunkten im jeweiligen Feld und die gemeinsame Akquise von weiteren europäischen Industrie- und Wissenschaftspartnern zur Stärkung der Innovation Community.

Industriepartner haben grundsätzlich die *volle Kontrolle* über die zur Verfügung gestellten *Daten*. Dies schließt sowohl die Erlaubnis von Zugriffen durch individuelle Forschungspartner ein als auch die Definition eines Lifecycle, der bspw. auch das Löschen der Daten beinhaltet. Weiter möchte das Smart Data Innovation Lab auch kleinen und mittleren Unternehmen zusätzliche Chancen bieten. Falls Software-Angebote von diesen Unternehmen das Portfolio des SDIL stärken können, kann das SDIL im Gegenzug dabei helfen, das Angebot der Unternehmen bekannter zu machen.