Aufbrechen der Datensilos – Big Data

Forschungsfragen aus dem Bereich Industrial Analytics

Benjamin Klöpper, Jan Christoph Schlake

Industrial Software Systems
ABB AG Forschungszentrum Deutschand
Wallstadter Str. 59
68526 Ladenburng
benjamin.kloepper@de.abb.com
jan-christoph.schlake@de.abb.com

Abstract: Aufgrund isolierter Datenhaltung und –verarbeitung in Datensilos kann das volle Potential der Analyse historischer Daten im Umfeld der Prozessindustire oftmals nicht ausgenutzt werden. Big Data Technologien bieten Chancen, diese Situation zu verbessern. Gleichzeit stellen sich aber neue Herausforderungen, die den Einsatz von Big Data in diesem Umfeld behindern.

1 Kurzfassung des Vortrags

In einer typischen Prozessanlage findet sich in der Regel eine Vielzahl von *Datensilos*, gesammelte Daten werden in isolierten Informationssystemen gespeichert und verarbeitet. Ähnlich wie im Fall von funktionalen Silos [Ame88] wird durch die isolierte Speicherung und Verarbeitung der Daten das Lernen aus historischen Daten durch Vergleichen und Erkennen von Korrelation erschwert. In einem solchen Zustand ist Datenanalyse eine sehr mühselige Aufgabe und erfordert sowohl tiefgehendes Verständnis für die Domäne oder sogar die einzelne Anlage wie auch erhebliche Personen- und Zeitaufwände. Big Data Technologien versprechen die Möglichkeit, Lösungen für Industrial Analytics leichter bereitzustellen und flächendeckend auszurollen. Gleichzeitig stellt dieses Anwendungsszenario einige Herausforderungen an die Big Data Research Community.

Big Data Technologien werden entwickelt, um Daten in großen Mengen (high volume), mit hoher Geschwindigkeit (high velocity) oder sehr unterschiedlichen Formaten (variety) zu behandeln [MB12]. Eine vollständige Integration aller Daten einer Prozessanlage haben sicherlich diese Eigenschaft: historische Prozessdaten in unkomprimierter Form erreichen Terabyte an Datenvolumen, die Daten haben unterschiedlichste Formate wie Zeitreihendaten oder unstrukturierten Text (z.B.

Schichtbücher). Sollen die Daten direkt während der Ausführung analysiert werden, müssen Datenströme mit einer hohen Samplingrate unter Echzeitanforderungen bearbeitet werden. Eine Vielzahl von Software existiert bereits, um diesen Anforderungen zu genügen.

Die zentrale nicht-funktionale Anforderung von *Industrial Analytics* sind aber niedrige Investment- und Betriebskosten sowie Nutzbarkeit durch Nicht-Informatiker. Diese Anforderungen lassen sich mit den heutigen Big Data Technologien nur eingeschränkt erreichen. Insgesamt bestehen vor allem folgende Herausforderungen:

Geringste Kosten für die Datenspeicherung: Der Einsatz von Standardhardware zur Datenspeicherung stellt einen ersten Schritt zur kostengünstigen Speicherung von Daten dar. Allerdings werden nicht immer alle Daten in unkomprimierter Form für Berechnungen benötigt. Die Verwendung bspw. eines Clusters mit dem Hadoop File Systems [SK+10] verursacht hier wesentlich größere Kosten als die Speicherung der Daten auf Network Attached Storages (NAS). Das Konzept des Data Lakes ist für *Industrial Analytics* nicht wirtschaftlich. Es werden Konzepte benötigt, Daten auf unterschiedliche Speichertechnologien zu verteilen und diese je nach analytischem Bedarf transparent zwischen den einzelnen Speicherformen zu verschieben.

Unterstützung von Machine Learning Methoden: Frameworks wie Mahout [OA+11] oder mlib für Spark setzen zwar Machine Learning Algorithmen auf Big Data Technologien um. Jedoch ist der Satz an Methoden nicht so vollständig wie in vergleichbaren „Small Data“ Bibliotheken. Zusätzlich unterscheiden sich die Umsetzungen und Aufrufe teilweise erheblich, je nachdem ob eine Batch-, In-Memory oder graph-basierte Technologie eingesetzt wird. Dies bedeutet einen hohen Portierung und Re-Implementierungsaufwand für *Industrial Analytics* Projekte.

Einfache Datenvorverarbeitung und –bereinigung: Datenvorverarbeitung und -bereinigung sind oftmals die zeitlich größten Aufwände in einem Industrial Analytics Projekt. Hier fehlen sowohl wiederverwendbare Vorgehensmodelle als auch generische Tools, die den Domänenexperten ohne Data Mining-Expertise effektiv unterstützen

Datensicherheit: Prozess- und Produktionsdaten können geschäftskritische Informationen wie bspw. Produktionsmenge oder Rezepturen offenlegen. Daraus ergeben sich starke Anforderungen an die Datensicherheit. Cloud-deployments von Big Data Technologien, die günstige Unterhaltskosten bieten, sind hier kritisch zu betrachten. Um im industriellen Umfeld akzeptiert zu werden, sind überzeugende Sicherheitskonzepte erforderlich.

Literaturverzeichnis

[Am88] AME Study Group on Functional Organization: Organizational Renewal – Tearing Down the Functional Silos, AME Target, 4 – 16, 1988

[ME12] McAfee, A.; Brynjolfsson, Erik: Big Data: The Management Revolution, Harvard Business Review, October 2013

[SK+10] Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R.: The hadoop distributed file system. In IEEE 26th Symposium on Mass Storage Systems and Technologies (pp. 1-10), 2010
[OA+11] Anil, R.; Dunning, T.; Friedman, E.: Mahout in action. Manning, 2011