ExDRa

Exploratory Data Science over Raw Data

Programm / Ausschreibung	IKT der Zukunft, IKT der Zukunft, IKT der Zukunft - AS DE-AT Datenwirtschaft	**Status**	abgeschlossen
Projektstart	01.06.2019	**Projektende**	31.08.2022
Zeitraum	2019 - 2022	**Projektlaufzeit**	39 Monate
Keywords	Data Science;Explorative Analyse;Datenintegration;Datenorganisation und Wiederverwendung;Föderierte Ausführung;		

Projektbeschreibung

Anwendungen des maschinellen Lernens (ML) auf Basis großer Datenmengen werden zunehmend auch im Unternehmenskontext eingesetzt um Wertschöpfungsprozesse zu verbessern und Wettbewerbsvorteile zu erwirken. Im Gegensatz zu klassischen ML Problemen sind diese Fragestellungen oft unterspezifiziert, erlauben unterschiedliche Analyseverfahren und können eine Vielzahl heterogener, verteilter, oder beschränkt zugänglicher, Datenquellen verwenden. Entsprechend ist der typische Data Science Prozess in Unternehmen explorativ, d.h. Data Scientists stellen Hypothesen auf, integrieren die notwendigen Daten, führen unterschiedliche Analysen durch und suchen damit nach interessanten Mustern und Modellen. Da der Mehrwert im Vorfeld der Analyse unbekannt ist werden kaum Investitionen in die systematische Akquise, Integration und Vorbereitung der Daten getätigt. Dies führt zu Redundanzen manueller Arbeitsschritte sowie ineffizienter Verarbeitung. Weiterhin ist die zentrale Konsolidierung technisch und ökonomisch nicht immer sinnvoll oder unterliegt Zugangsbeschränkungen (z.B. sensible Daten). Diese Szenarien verbindet die Not-wendigkeit der föderierten Ausführung und der gezielten Redundanzeliminierung.

Die Idee des ExDRa Projekts ist es geeignete Systemunterstützung für diesen explorativen Data Science Prozess über heterogene und verteilte Rohdatenquellen zu untersuchen und im Rahmen eines Demonstrators für praktische Anwendungen bereitzustellen. Im Detail umfasst der Ansatz die Forschungsschwerpunkte (1) ad-hoc und föderierte Datenintegration über Rohdaten, (2) Datenorganisation und Wiederverwendung von Zwischenergebnissen, (3) horizontale Optimierungen über den gesamten Data Science Lebenszyklus, und (4) Anfrageplanung für beschränkt zugängliche Datenbestände. Als Anwendungsfall dient die Prozessindustrie (z.B. Chemie, Pharma, Wasser, Öl und Gas) bei der Siemens AG. In diesem Kontext existieren große Datenmengen, welche über Standorte und Anlagen verteilt sind, und deren Konsolidierung technisch, ökonomisch, und rechtlich eingeschränkt ist.

Aus dem Gesamtziel resultieren vier Arbeitsziele. Erstens ist die Datenintegration, Datenvorbereitung, und Analyse von Rohdaten, mittels einer geeigneten deklarativen Beschreibung von Datenquellen und Vorverarbeitungsschritten sowie effizienter Primitive der lokalen und föderierten Ausführung, zu ermöglichen. Im Kontext explorativer Data Science erfordert dies geeignete Stichprobenverfahren und Techniken der inkrementellen Wartung. Zweitens, sind unnötige Redundanzen und

Ineffizienzen wiederholter Verarbeitungsschritte durch Methoden der Datenorg-anisation und Wiederverwendung zu beheben. Der hohe Kommunikationsaufwand föderierter Analysen erfordert weiterhin eine Untersuchung von Kompressions-techniken und des Performance-Genauigkeits-Tradeoffs. Drittens, soll mit Hilfe einer systematischen Modellverwaltung und Optimierung von Experimenten die Nachvollziehbarkeit von explorativen Analyseergebnissen verbessert und zukünftige Analysen erleichtert werden. Viertens, ist die föderierte Verarbeitung ein integraler Bestandteil der explorativen Analyse von beschränkt-zugängliche Rohdaten. Hier sollen geeignete Systemarchitekturen und Methoden der Anfrageplanung und -ausführung untersucht werden. Um die praktische Anwendbarkeit nachzuweisen werden die Ergebnisse in eine Demonstrator-Software integriert und erprobt.

Abstract

Machine learning (ML) applications based on large data are increasing applied in the enterprise to improve the value chain and gain competitive advantage. In contrast to traditional ML, the objectives are, however, under-specified, allow for different types of analysis, and can leverage a wide variety of heterogeneous, distributed and partially inaccessible data sources. Therefore, the typical data science process in the enterprise is exploratory, that is, data scientists investigate hypotheses, integrate the necessary data, run different analytics, and look for interesting patterns and models. Since the added value is unknown in advance, very little investments are made into the systematic acquisition, integration, and preprocessing of data. This lack of infrastructure results in redundancy of manual steps and inefficient computation. Furthermore, the central consolidation is not always technically or economically desirable or even possible (e.g., sensitive personal data). These scenarios share the necessity of federated execution and dedicated elimination of redundancy.

The basic idea of the ExDRa project is an investigation of suitable systems support for this exploratory data science process over heterogeneous and distributed raw data sources, showcased in a demonstrator for practical applications. In detail, this approach entails the following research aspects: (1) ad-hoc and federated data integration over raw data, (2) data organization and reuse of intermediates, (3) horizontal optimization over the entire data science lifecycle, and (4) query planning for partially accessible data. Use cases come from the process industry at Siemens AG. In this context, there are large amounts of data, distributed over locations and appliances, but whose consolidation is technically, economically, and legally limited.

The overall goal leads to four research goals. First, data integration, data processing and analysis over raw data needs to be enabled via a suitable declarative specification of data source and preprocessing steps, as well as efficient primitives for local and federated computation. In the context of exploratory data science, this requires sampling and incremental maintenance. Second, unnecessary redundancy and inefficiency of repeated computations need to be addressed via dedicated techniques for data organization and reuse. The high communication overhead of federated analysis could further benefit from leveraging compression techniques and the performance-accuracy tradeoff. Third, we aim to improve the understanding of exploratory analysis results and simply future analysis via systematic model management and optimization of experiments. Fourth, federated computation is an essential part of exploratory analysis over raw data. Accordingly, we intend to investigate system architectures, as well as query optimization and processing. In order to provide evidence for practical relevance, all results will be integrated and evaluated as part of a demonstrator software.

Projektpartner

- Technische Universität Graz