Big Data - Datenquellen und Anwendungen

AW1 Präsentation

Gerrit Thede

Fakultät Technik und Informatik
Department Informatik
HAW Hamburg

18. November 2013

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Was ist „Big“ Data?

Herausforderungen

- **Volume**: Menge an Daten
- **Velocity**: Geschwindigkeit der Verarbeitung und Entscheidungsfindung
- **Variety**: unstrukturierte komplexe Daten

Was bedeutet „Big“ Data?

Smart Data erschaffen

- Analyse der Daten
- Kombination mit anderen Datenquellen
- Prozess der Informationsgewinnung

Qualitativ hochwertige Datenquellen

Es gibt mehr Big Data als Twitter & Co.!

Datenquellen anzapfen

- Sensordaten
- Mobilitätsdaten aus Telefonen, Geodaten
- Quantified Self
- Logfiles
- Social web
- Crowd Sourcing, Communities
- Open Data
- Datenhändler

Data Scientists

Neue Berufsgruppen

Benötigte Fähigkeiten

- Mathematik
- Statistik
- Maschinelles Lernen
- Natural Language Processing
- mit Daten spielen

Data Science Community

Plattform Kaggle

- Plattform für Firmen, die Problemlösungen suchen
- Preisgeld für die beste Datenanalyse
- Bereitstellung von Datensets
- Wettbewerb für Data Scientists
- Beispiel: General Electric, „Flight Quest 2: Flight Optimization Optimize flight routes based on current weather and traffic.“ Preisgeld: USD 250.000

http://www.kaggle.com

Data Journalism

Daten Visualisierung

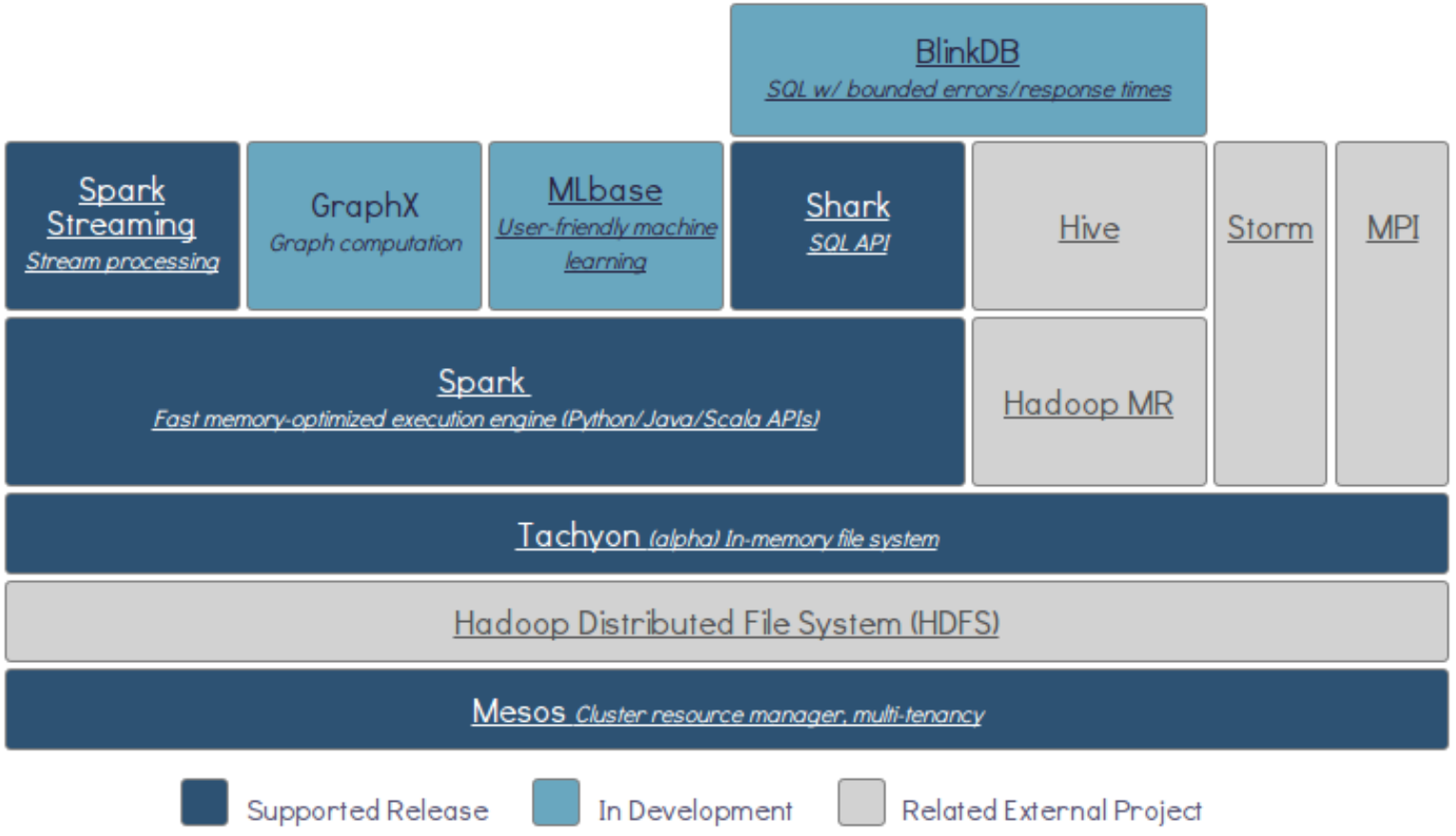
- attraktiv aufbereitet
- verständlich
- Zusammenhänge erkennbar in Grafiken, nicht in Zahlenreihen
- Geovisualisierung
- Data Storytelling

Open Data
Open Government

Data is the new raw material of the 21st century.
(Nigel Shadbolt, Tim Berners-Lee)

- Verwaltungen werden verpflichtet, Daten den Bürgern zur Verfügung zu stellen
- Beispiel: Spesenabrechnungen der britischen Regierung, 700.000 Dokumente mit Einzelbelegen, Analyse schwierig, erfolgreich mit Crowdsourcing durch Guardian. Ergebnis: 27000 Benutzer überprüften 460000 Dokumente und führte zu Überprüfung der Abrechnungen.

Technologie
BDAS, the Berkeley Data Analytics Stack



Berkeley Data Analytics Stack [https://amplab.cs.berkeley.edu/software/]

Berkeley Data Analytics Stack

Alternative zu Hadoop

Apache Spark (Incubator Project seit Juni 2013)

- Datenanalyse Cluster Computing Framework
- Aufbau auf Hadoop File System (HDFS)
- im Gegensatz zu Hadoop mit „In-Memory“ Technik
- verspricht, bis zu 100 mal schneller als Hadoop zu sein
- Stapelverarbeitung und Streaming Analyse
- interaktive Analyse (Cachen von Daten, Anwenden verschiedener Berechnungen)
- Programmierung mit Scala, Python statt PIG
- Framework für maschinelles Lernen
- kompatibel und interoperabel mit Hadoop Ökosystem

Personen: Ion Stoica, Matei Zaharia (UC Berkeley)

Predictive Analytics

Bausteine

- historische Daten und Fluss neuer Daten
- Statistik
- Modellbildung
- Maschinelles Lernen
- Data Mining

Ziel ist die Vorhersage von Fragestellungen zur Entscheidungsfindung.

FuturICT

- Dirk Helbing, ETH Zürich
- Plattform für Simulation, Visualisierung und Partizipation von sozialen Systemen
- Entwicklung einer Spiegelwelt zur Beantwortung von „Was wäre, wenn ...?“ Fragen.
- Proposal für EU Flagship Project 2013 (10 Jahre, 1 Milliarde €)

http://www.futurict.eu

FuturICT

FuturICT

- Planetary Nervous System:
  - weltweite Sensordaten
  - füttert die Modelle mit Daten
- Living Earth Simulator
  - Modelle für Gesundheit, Energie, Infrastruktur, Verkehr, Finanzen, ...
- Global Participatory Platform
  - Offenes Framework für Bürger, Firmen und Organisationen
  - Übermittlung von neuen Daten und erforschen von Daten

Big Data bei Versicherungen

Möglichkeiten für Krankenversicherung

- aus Krankenhausabrechnungen Vorhersage für zukünftige Krankenhausaufenthalte
- vorsorgliche Präventivleistungen
- Betrugserkennung
- unrentable Vertragsverlängerungen

- Interactive Learning for Efficiently Detecting Errors in Insurance Claims [GK11]
- Knowledge Discovery from Massive Healthcare Claims Data [CSS13]

Konferenzen

- ACM SIGKDD Knowledge Discovery and Data Mining [http://www.kdd.org/]
- ACM SIGMOD Management of Data [http://www.sigmod.org]

Personen

- Alex Pentland (Direktor des MIT Human Dynamics Laboratory und des MIT Media Lab), Reality Mining, Personal Data

New Deal on Data

1. Recht, Daten zu besitzen. Firmen sollen Daten nur verwalten und entfernen wenn Nutzer es will (wie Schweizer Bankkonto).
2. Der Datenbesitzer muss die volle Kontrolle über die Daten haben. Möglichkeit darüber zu bestimmen, wie die Firma die Daten verarbeitet. Nur Opt-In.
3. Recht, die Daten wiederzuerlangen, zu löschen oder woanders zu verwenden.

[Pen09], World Economic Forum

Ziel ist die Nutzung von Big Data im Einklang mit der Privatsphäre, um den positiven Nutzen der Allgemeinheit zu ermöglichen.

Ende

Vielen Dank für Ihre Aufmerksamkeit!

Referenzen

[CSS13] Chandola, Varun, Sreenivas R. Sukumar und Jack C. Schryver: *Knowledge Discovery from Massive Healthcare Claims Data*. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, Seiten 1312–1320, New York, NY, USA, 2013. ACM.

[GK11] Ghani, Rayid und Mohit Kumar: *Interactive Learning for Efficiently Detecting Errors in Insurance Claims*. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, Seiten 325–333, New York, NY, USA, 2011. ACM.

[Pen09] Pentland, Alex (MIT): *Reality Mining of Mobile Communications: Toward a New Deal on Data*. The Global Information Technology Report 2008-2009 World Economic Forum, 2009.