The Berlin Big Data Center

Jonas Traub
Technische Universität Berlin / DFKI IAM
www.dima.tu-berlin.de | bbdc.berlin | jonas.traub@tu-berlin.de

with materials from Tilmann Rabl and Volker Markl
BBDC Goals

• **Pooling expertise** in scalable data management, data analytics, and big data applications.

• **Conducting fundamental research** to develop novel and automatically scalable technologies capable of performing “deep analysis” of big data.

• **Developing an integrated, declarative, highly scalable, open-source system** that enables the specification, automatic optimization, parallelization, hardware adaptation, fault-tolerance, and efficient execution of advanced data analysis problems using varying methods that leverage our work on Apache Flink.

• **Transferring technology and know-how** to support innovation in companies and startups.

• **Educating data scientists** with respect to the five big data dimensions via leading educational programs.

• **Empowering people to leverage** “Smart Data.”

• **Enabling the general public to conduct sound data-driven decision-making.**
Big Data – System View

Tension between *performance* and *algorithmic expressiveness*
“Data Scientist” – “Jack of All Trades!”

Domain Expertise (e.g., Industry 4.0, Medicine, Physics, Engineering, Energy, Logistics)
Mathematical Programming
Linear Algebra
Stochastic Gradient Descent
Error Estimation
Active Sampling
Regression
Monte Carlo
Statistics
Sketches
Hashing
Convergence
Decoupling
Iterative Algorithms
Curse of Dimensionality

Relational Algebra / SQL
Data Warehouse/OLAP
NF²/XQuery
Resource Management
Hardware Adaptation
Fault Tolerance
Memory Management
Parallelization
Scalability
Memory Hierarchy
Data Analysis Language
Compiler
Query Optimization
Indexing
Data Flow
Control Flow
Real-Time

Data Science
Data Analysis
Scalable Data Management
Application
Big Data Analytics Requires Systems Programming

Data Analysis
Statistics
Algebra
Optimization
Machine Learning
NLP
Signal Processing
Image Analysis
Audio-, Video Analysis
Information Integration
Information Extraction
Data Value Chain
Data Analysis Process
Predictive Analytics

R/Matlab: 5 million users

Hadoop: 200,000 users

“We will soon have a huge skills shortage for data-related jobs.”
Neelie Kroes (ICT 2013, Nov. 7, Vilnius)

People with Big Data Analytics Skills
Indexing
Parallelization
Communication
Memory Management
Query Optimization
Efficient Algorithms
Resource Management
Fault Tolerance
Numerical Stability

“Big Data is now where database systems were in the 70s (prior to relational algebra, query optimization and a SQL-standard)!

Declarative languages to the rescue!
“Data Scientist” – “Jack of All Trades!”

Domain Expertise (e.g., Industry 4.0, Medicine, Physics, Engineering, Energy, Logistics)
Mathematical Programming
Linear Algebra
Stochastic Gradient Descent
Error Estimation
Active Sampling
Regression
Monte Carlo
Statistics
Sketches
Hashing
Convergence
Decoupling
Iterative Algorithms
Curse of Dimensionality

Data Science
Statistics
Machine Learning

Danger Zone!

Traditional Research

Hacking Skills

New Technology to the Rescue!

Relational Algebra / SQL
Data Warehouse/OLAP
NF²/XQuery
Resource Management
Hardware Adaptation
Fault Tolerance
Memory Management
Parallelization
Scalability
Memory Hierarchy
Data Analysis Language
Compiler
Query Optimization
Indexing
Data Flow
Control Flow
Real-Time

© DIMA 2017
Danger Zone!

Total revenue generated by arcades correlates with Computer science doctorates awarded in the US
Danger Zone! Contd.

Divorce rate in Maine correlates with Per capita consumption of margarine
Apache Flink—
A Success Story
Born in Berlin

http://flink.apache.org
Timeline

- 2008: Initial Vision for a Big Data Analytics Platform
- 2009: DFG Proposal for StratoSphere I
- 2010: Grant Award Start of StratoSphere I
- 2010: Nephele/PACTs Paper Published
- 2011: Flink Forward Conference
- 2012: Grant Award Start of StratoSphere II
- 2012: Spinning Fast Iterative Dataflows Paper Published
- 2014: Apache Flink Incubator Project
- 2014: Stratosphere System Paper Published
- 2014: The VLDB Journal
- 2014: Berlin Big Data Center Founded
- 2014: Apache Flink Top Level Project
- 2015: dataArtisans Founded
- 2015: FlinkForward Conference in Berlin
- 2016: FlinkForward Conference in San Francisco
- 2016: Flink Community Groups Across Europe
- 2016: Berlin 830
- 2016: Paris 520
- 2016: Madrid 457
- 2016: Stockholm 337
- 2016: Brussels 325
- 2016: London 315
- 2016: Munich 171
- 2016: Amsterdam 124
- 2016: Istanbul 67

21 Meetups Worldwide
250 Contributors
516 Members
26 Cities
14 Countries
What can I do with it?

A big data processing system that can **natively** support all these workloads.

- **Stream processing**
  - Delta of 5% price
  - Warning
  - Count
  - Tumbling 30 sec window
  - Sum

- **Batch processing**

- **Machine Learning at scale**

- **Graph Analysis**
What is Apache Flink?

Apache Flink® is an open-source stream processing framework for distributed, high-performing, always-available, and accurate data streaming applications.

**Key Features**

- Bounded and unbounded data
- Event time semantics
- Stateful and fault-tolerant
- Running on thousands of nodes with very good throughput and latency
- Exactly-once semantics for stateful computations
- Flexible windowing based on time, count, or sessions in addition to data-driven windows
- **DataSet** and **DataStream** programming abstractions are the foundation for user programs and higher layers

---

Companies Using Flink

[Logos of companies using Flink]
Innovation und Transfer am Beispiel Apache Flink

Flink Community

17.820+ Meetup-Mitglieder weltweit
328+ open-source Entwickler (contributors)
40+ Meetupgruppen weltweit

14+ Länder mit regelmäßigen Meetup-Events
30+ Anwenderunternehmen
Firmengründung dataArtisans
Machine Learning + Data Management = X

Think ML-algorithms in a scalable way

Process iterative algorithms in a scalable way

Goal: Data Analysis without System Programming!

Feature Engineering
Representation Algorithms (SVM, GPs, etc.)

Mathematical Programming
Linear Algebra
Error Estimation
Active Sampling
Regression Monte Carlo

Statistic
Sketches
Hashing
Isolation
Convergence
Curse of Dimensionality
Iterative Algorithms
Control flow

Relational Algebra/SQL
Data Warehouse/OLAP
NF²/XQuery
Scalability
Hardware adaption
Fault Tolerance
Resource Management

Declarative Languages
Automatic Adaption
Scalable processing

Parallelization
Compiler
Memory Management
Memory Hierarchy
Data Analysis Language
Query Optimization
Dataflow
Indexing

© DIMA 2017
What, Not How! Consider K-means Clustering.

Declarative data analysis program with automatic optimization, parallelization and hardware adaption

65 lines of code
short development time
robust runtime

"What"

Apache Flink (Scala frontend)

Hand-optimized code
(data-, load- and system dependent)

486 lines of code
long development time
non-robust runtime

"How"
Big Data Analytics Without Systems Programming! (What, Not How!)

Description of "What?" (declarative specification)

- Larger human base of "data scientists"
- Reduction of "human" latencies
- Cost reduction

Description of "How?" (state of the art in scalable data analysis)
- Map/Reduce, MPI

Data Analyst

Machine
Flink in the BBDC Stack
BBDC Lessons Learned & Challenges

• International competitors have far more funding and visibility
  – Berkeley AmpLab: $30 Mio for 6 years, recently grown into Berkeley Institute for Data Science with another $40 Mio funding
  – UK Turing Institute: GBP 67 Mio funding
  – And many others (e.g., across the US, China, Korea, and Japan)

• German companies are followers, not leaders in big data
  – Many of Germany’s large companies have not yet developed a big data strategy and are risk-averse, or focus too much on short-term and established solutions.
  – It is far easier to work with “new“ companies to transfer novel technologies
    • ResearchGate, Zalando, King, IMR, and Spotify, among others
  – Open source solutions and/or establishing new companies are the best route to turn research into innovation
  – US and large international companies are easier to collaborate with, often times via their respective German subsidiary.
Thank You

Contact:
Jonas Traub
jonas.traub@tu-berlin.de
The Berlin Big Data Center

Jonas Traub
Technische Universität Berlin / DFKI IAM
www.dima.tu-berlin.de | bbdc.berlin | jonas.traub@tu-berlin.de

with materials from Tilmann Rabl and Volker Markl