



Data Analysis with Spark — from Raw Data to Insights

Prof. Dr. Markus Breunig
Stephan Schiffner



<Short Film introducing the Lab- Automation>

Commercial Objectives

Throughput Optimization

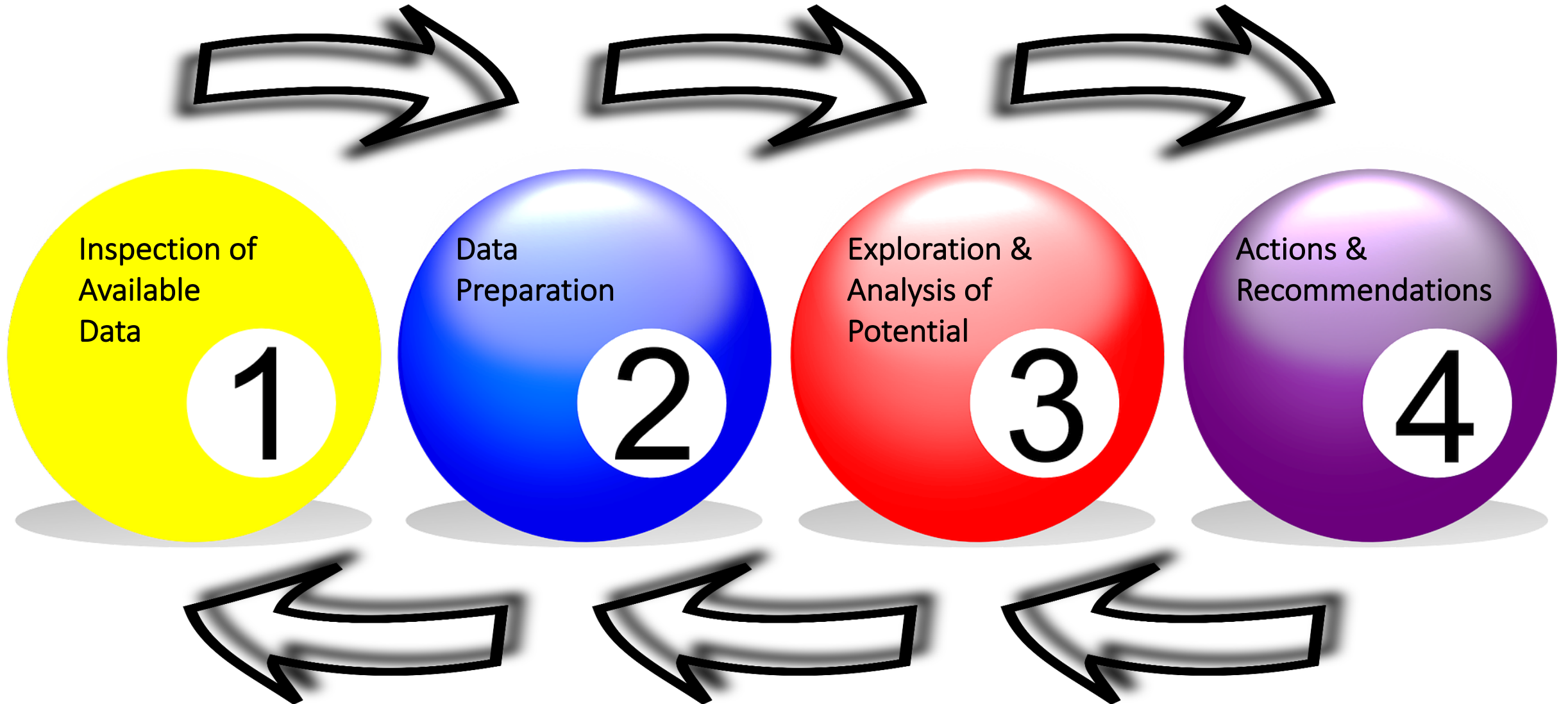
→ Fully analyze as many samples as possible per day

Cycle Time Optimization

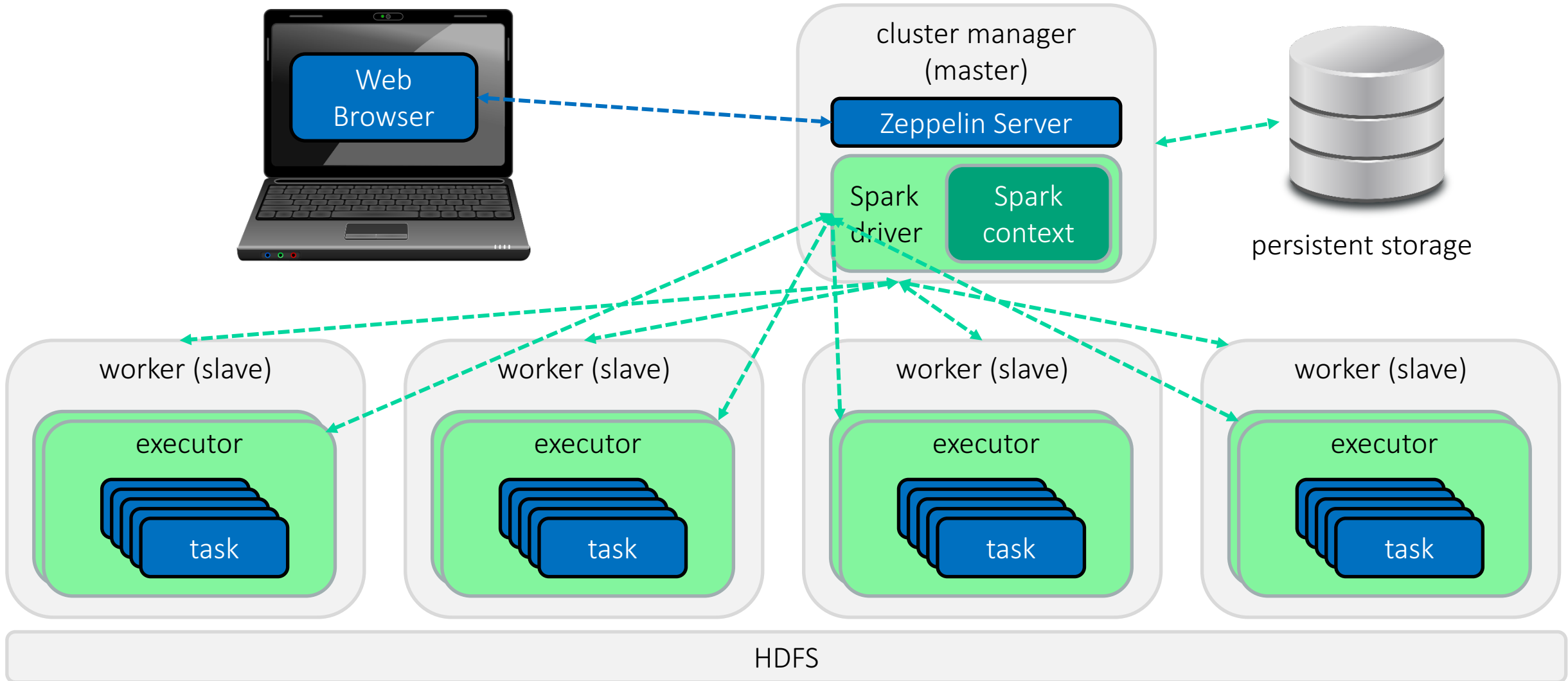
→ Reduce the max/avg time each samples needs to be fully analyzed

Physical Constraint: extension of the track very expensive/impossible

Idea: identify optimization potentials based on the available log-data generated by the „track sample manager“ software



Log-Data Analysis (performed on a Spark Cluster)

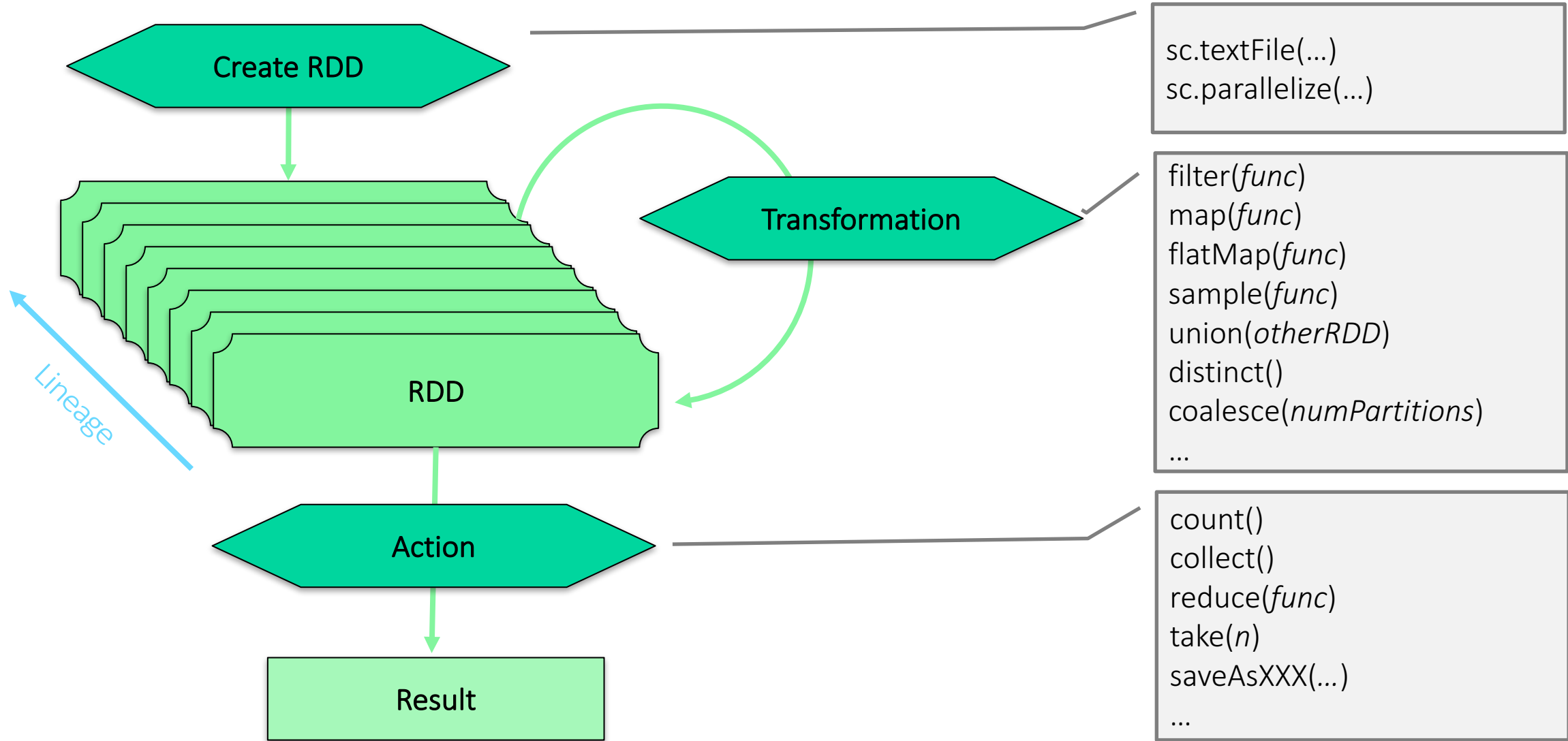




Part I

Data Inspection:
Import of the Log-Files

RDD – Resilient Distributed Dataset





Part II

Data Preparation and
Data Exploration and
Visualization

- Data Analytics Process
 - From the Business Objective
 - via Insights gleaned from the Data
 - to Recommendations and Actions
- Data Analytic Tools Used
 - Spark and Spark Data Structures (RDDs)
 - SQL and Visualizations
- Outlook
 - Complex Predictive Analysis (MLlib, ...)



Contact Persons at this Conference



Stephan Schiffner
Head of
Big Data & Analytics



Prof. Dr. Markus Breunig
Lead Consultant
Big Data & Analytics



Oliver Hold
Head of Sales
Solutions & Consulting



Dr. Federica Fusco
Data Scientist
Big Data & Analytics

*Workshop (tomorrow):
TensorFlow step-by-step*

Come and Visit us at our Exhibition Booth!