Funded by
the European Union

ASSISTANT

Evaldas Vaičiukynas (evaldas.vaiciukynas@ktu.lt)

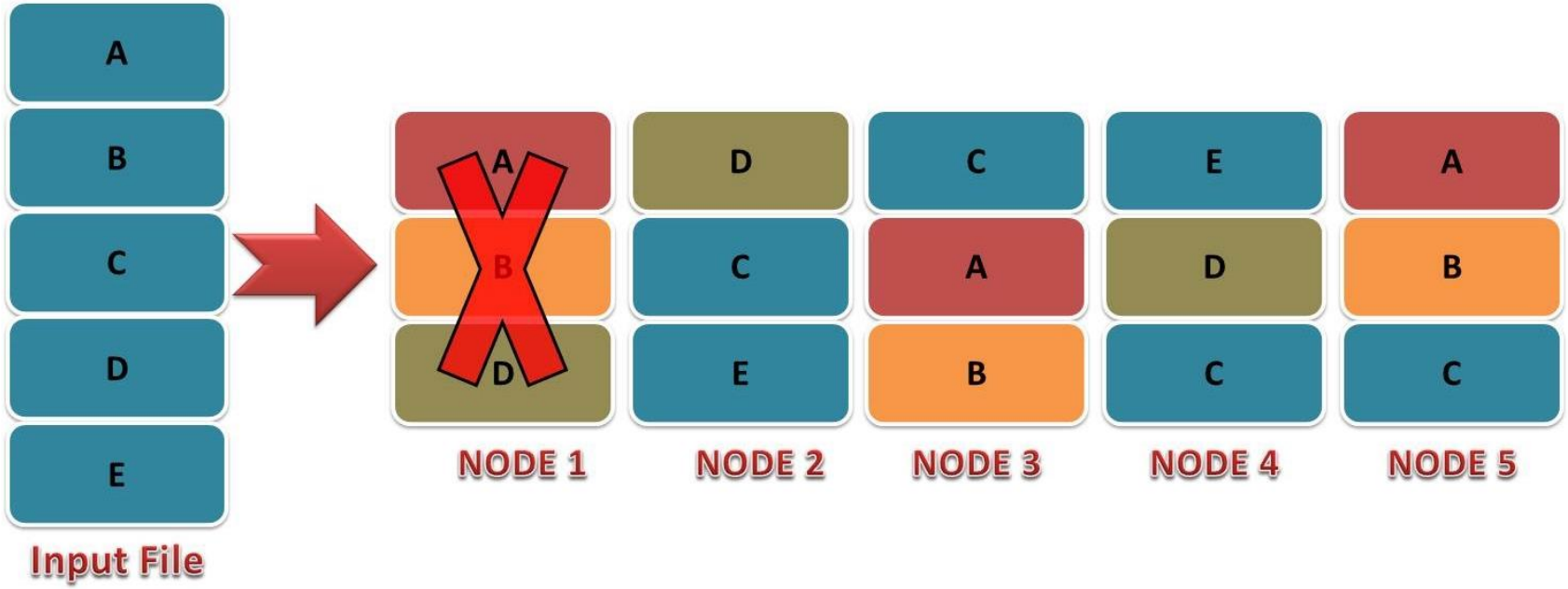
https://www.assistant-erasmus.eu

Tools for Big Data Analytics

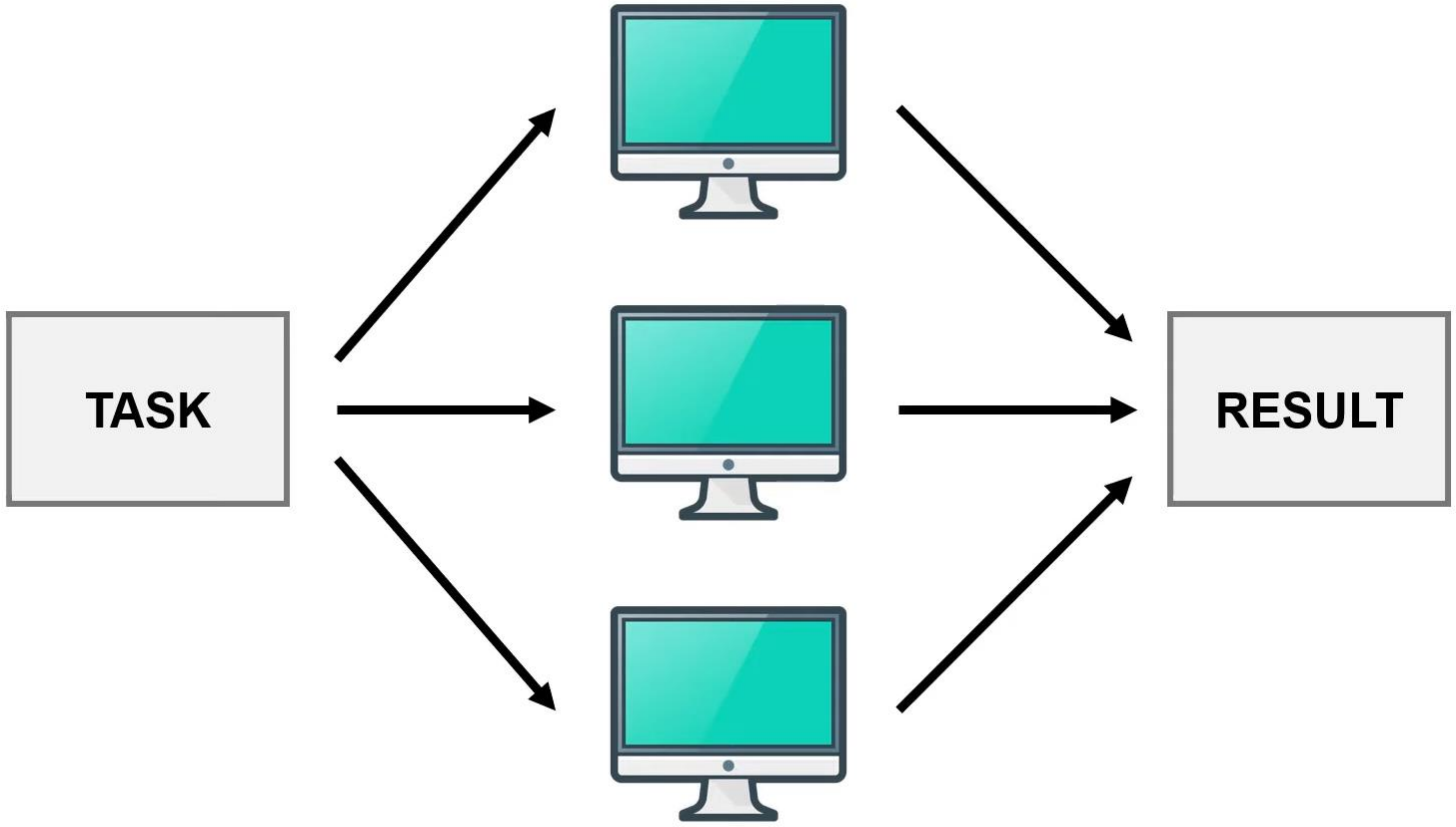
Hadoop Distributed File System (HDFS) = Fault Tolerance

- File Size = 640 MB
- Block Size = 128 MB
- Replication Factor = 3

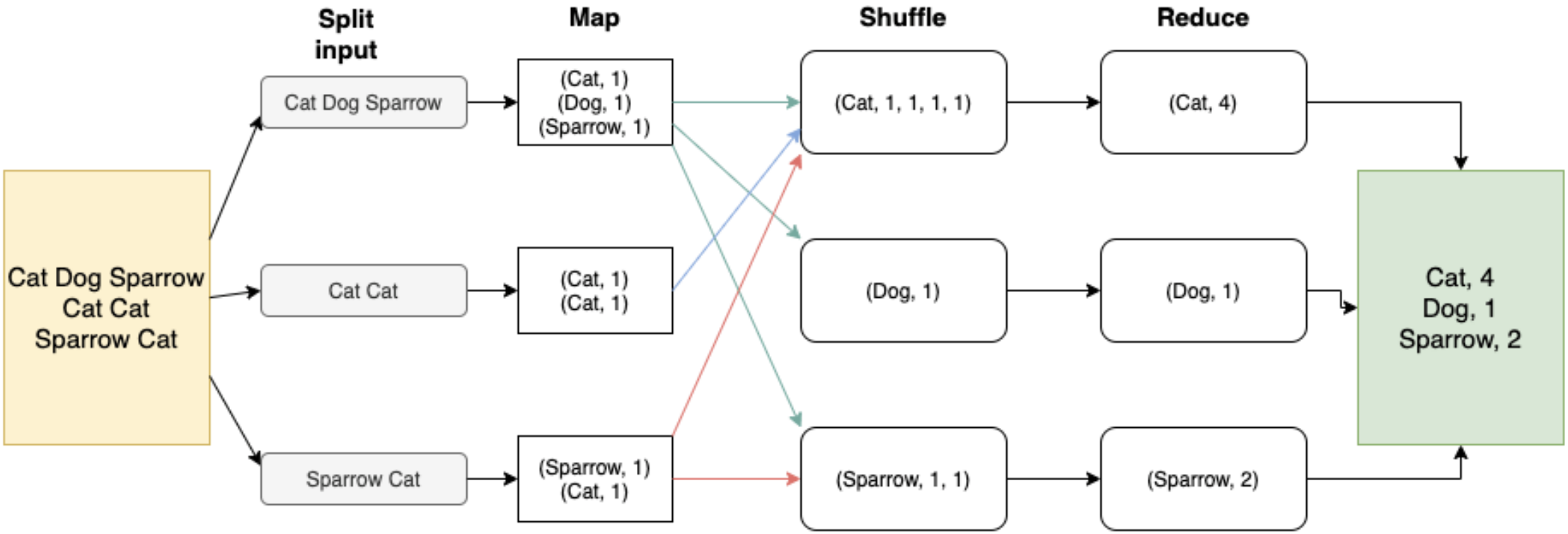
  200% Disk Overhead:
  384 MB x 5 = 1920 MB



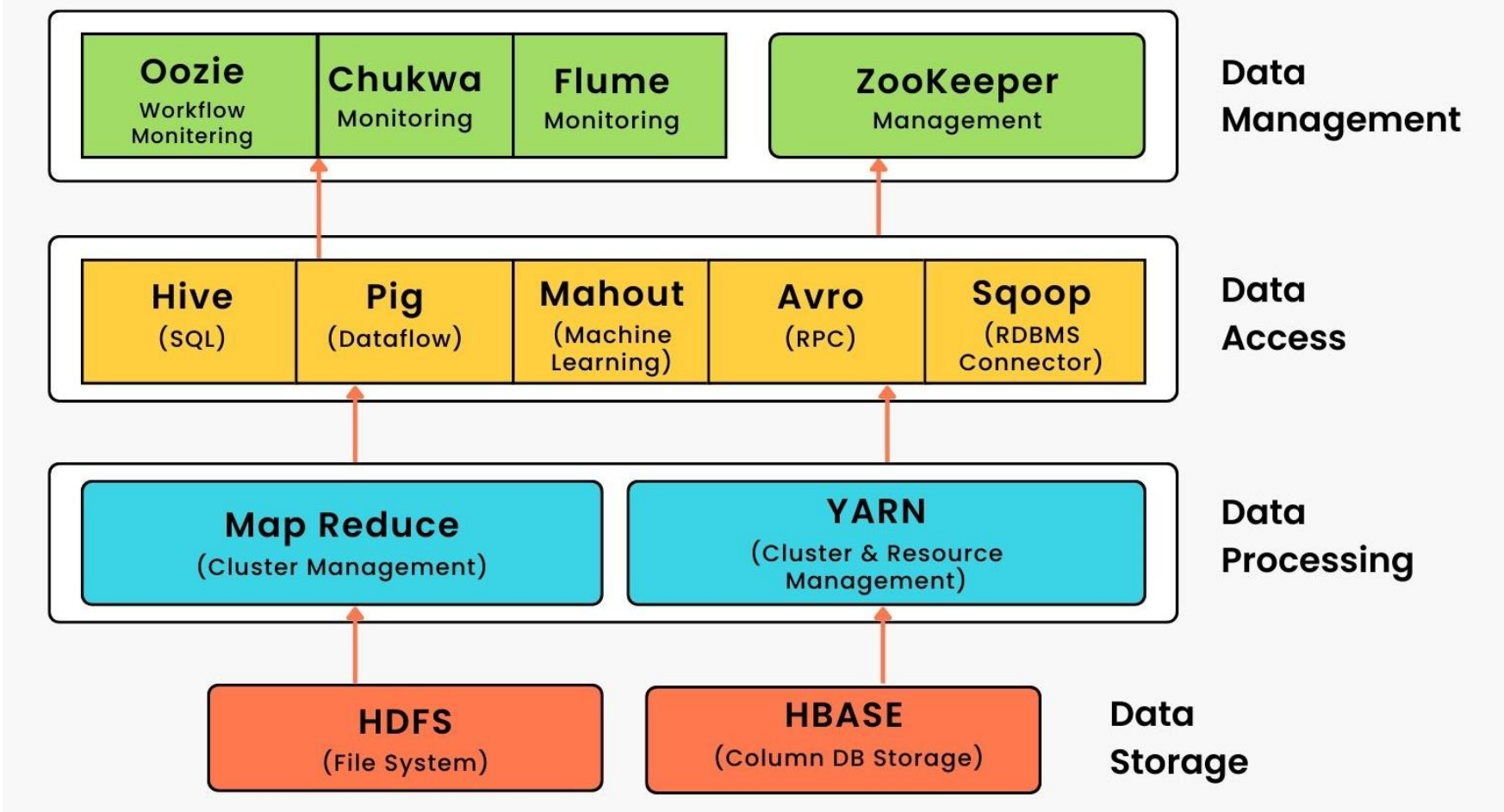
Main Idea: Parallel Computing



MapReduce = split + apply + combine (fold)



Hadoop Ecosystem



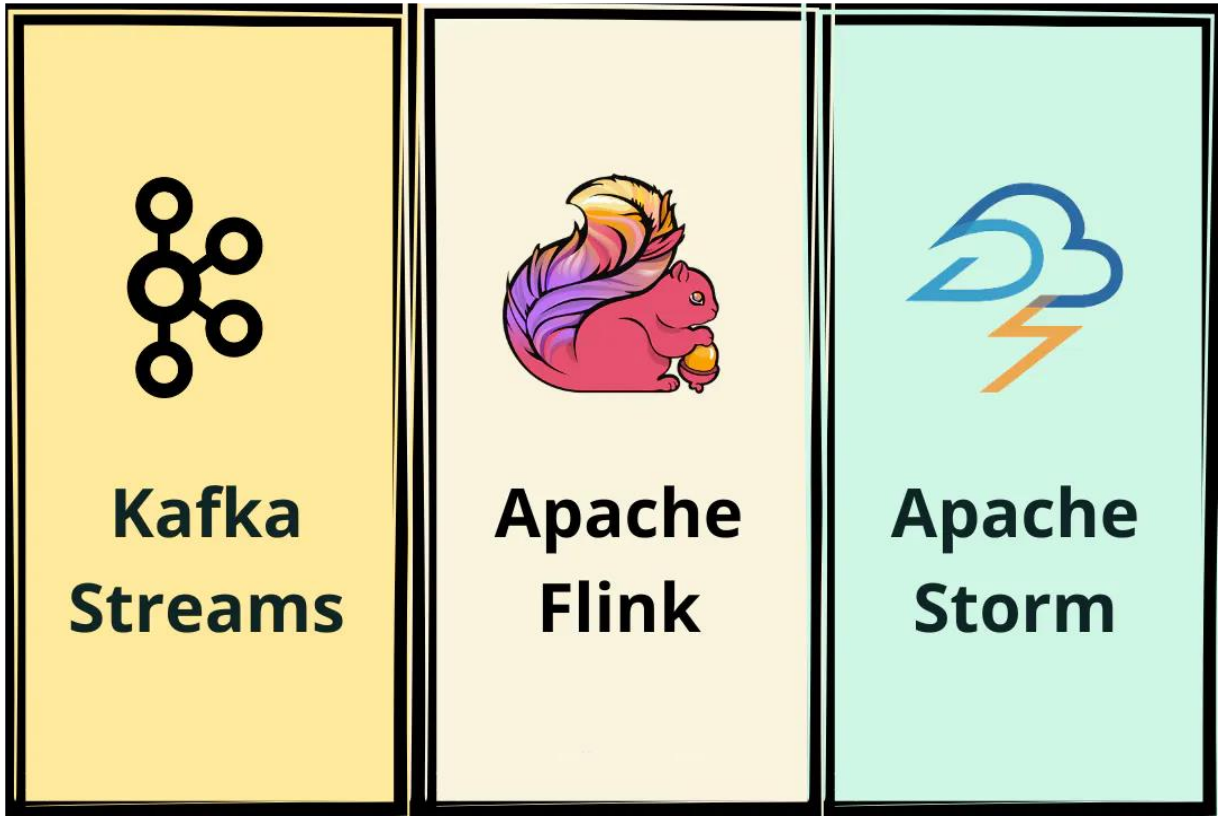
Tools for Manipulating Big Data (2014)

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28. https://doi.org/10.1257/jep.28.2.3

Google name	Analog	Description
Google File System	Hadoop File System	This system supports files so large that they must be distributed across hundreds or even thousands of computers.
Bigtable	Cassandra	This is a table of data that lives in the Google File System. It too can stretch over many computers. It was created by Facebook back in 2008.
MapReduce	Hadoop	This is a system for accessing and manipulating data in large data structures such as Bigtables. MapReduce allows you to access the data in parallel, using hundreds or thousands of machines to extract the data you are interested in. The query is “mapped” to the machines and is then applied in parallel to different shards of the data. The partial calculations are then combined (“reduced”) to create the summary table you are interested in.
Sawzall	Pig	This is a language for creating MapReduce jobs.
Dremel, BigQuery	Hive, Drill, Impala	This is a tool that allows data queries to be written in a simplified form of of Structured Query Language (SQL). With Dremel it is possible to run an SQL query on a petabtye of data (1,000 terabytes) in a few seconds.

Solutions for Streaming Data

- Apache Kafka Streams – library for building apps and microservices through producer-consumer architecture. Languages: Java, Scala.
- Apache Flink – stream processing framework for high-performance, high-throughput, fault-tolerant, and real-time analytics. Languages: Java, Scala, Python, SQL.
- Apache Storm – distributed real-time computation system, designed to process large volumes of high-velocity data. Useful for real-time analytics, online machine learning, continuous computation, distributed RPC, ETL, etc. Languages: JVM, Ruby, Python, Javascript.



H2O: distributed in-memory machine learning platform

H2O: supported machine learning algorithms

- H2O supports the following supervised algorithms:
  - Cox Proportional Hazards (CoxPH)
  - Deep Learning (Neural Networks)
  - Distributed Random Forest (DRF)
  - Generalized Linear Model (GLM)
  - Isotonic Regression
  - Generalized Additive Models (GAM)
  - ANOVA GLM
  - Gradient Boosting Machine (GBM)
  - Naïve Bayes Classifier
  - Decision Tree
  - AdaBoost
  - Stacked Ensembles, RuleFit
  - Support Vector Machine (SVM)
  - Distributed Uplift Random Forest (Uplift DRF)
  - XGBoost

- H2O supported languages: Python, R, Java, and Scala.

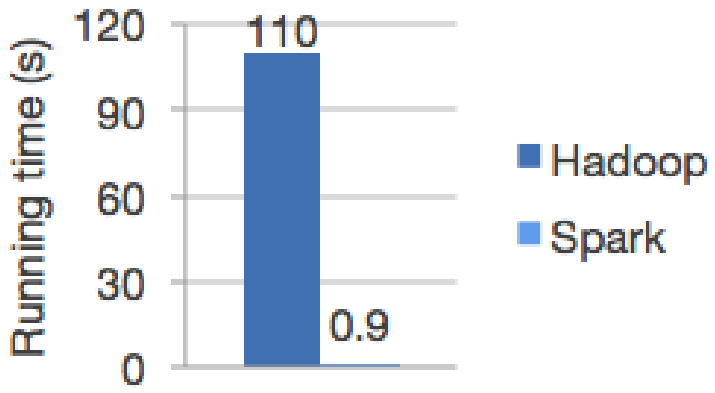
- H2O supports the following unsupervised algorithms:
  - Aggregator
  - Generalized Low Rank Models (GLRM)
  - Isolation Forest
  - Extended Isolation Forest
  - K-Means Clustering
  - Principal Component Analysis (PCA)

Apache Spark MLlib: scalable machine learning library

- MLlib supports these languages: Java, Scala, Python, and R.

- ML algorithms include:
  - Classification: logistic regression, naive Bayes,...
  - Regression: generalized linear regression, survival regression,...
  - Decision trees, random forests, and gradient-boosted trees
  - Recommendation: alternating least squares (ALS)
  - Clustering: K-means, Gaussian mixtures (GMMs),...
  - Topic modeling: latent Dirichlet allocation (LDA)
  - Frequent itemsets, association rules, and sequential pattern mining

Performance
High-quality algorithms,
100x faster than MapReduce.



Thank you for attention

Funded by
the European Union