rapidminer

#1 Modern Platform to
Turn Data into a Strategic Asset

Best Practices for Big Data Analytics in Hadoop

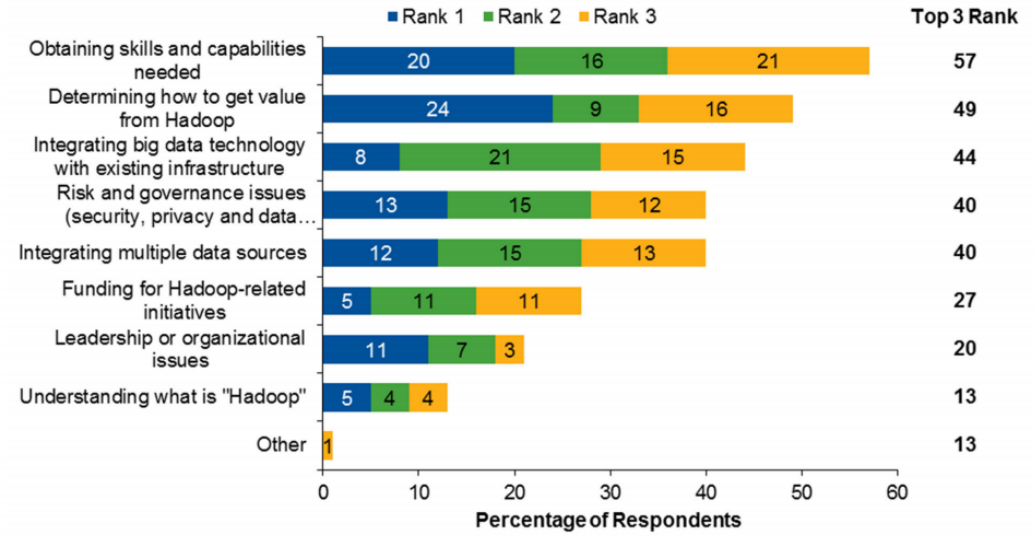
Zoltan Prekopcsak
VP Big Data

June 14, 2016

90% of Deployed Data Lakes are "USELESS"

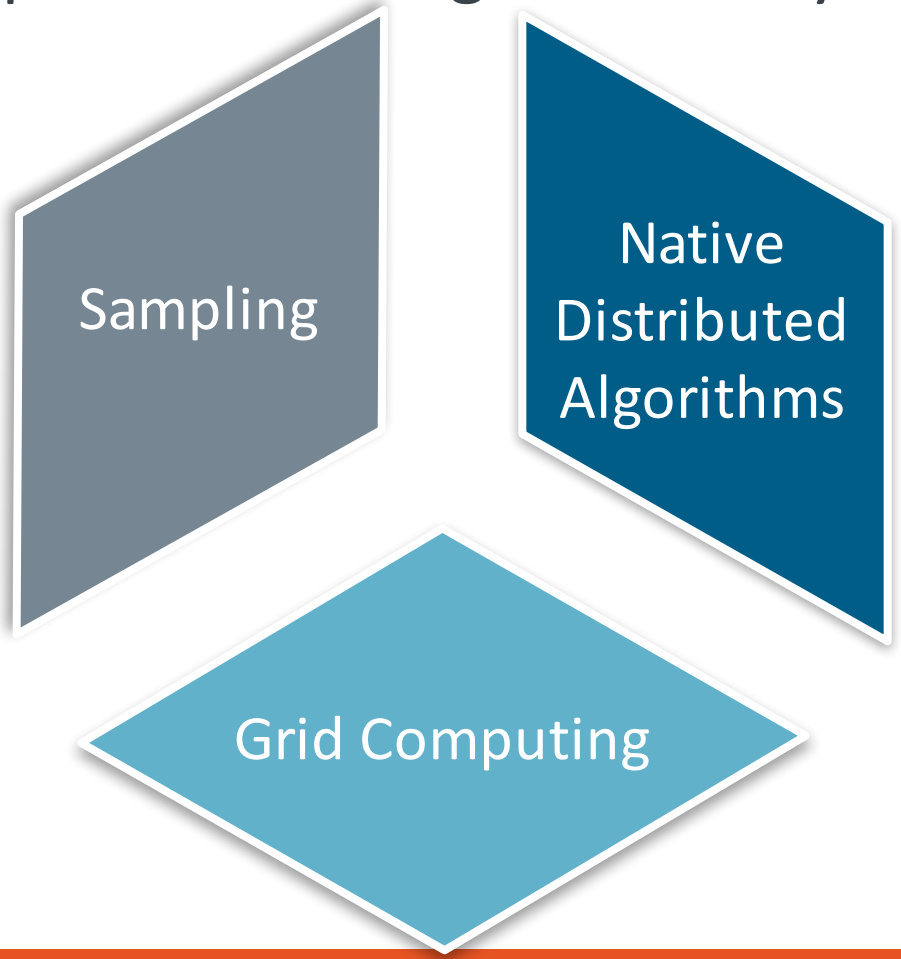
"Through 2018, 90% of deployed data lakes will be useless as they are overwhelmed with information assets captured for uncertain use cases."

Figure 5. Hadoop Challenges

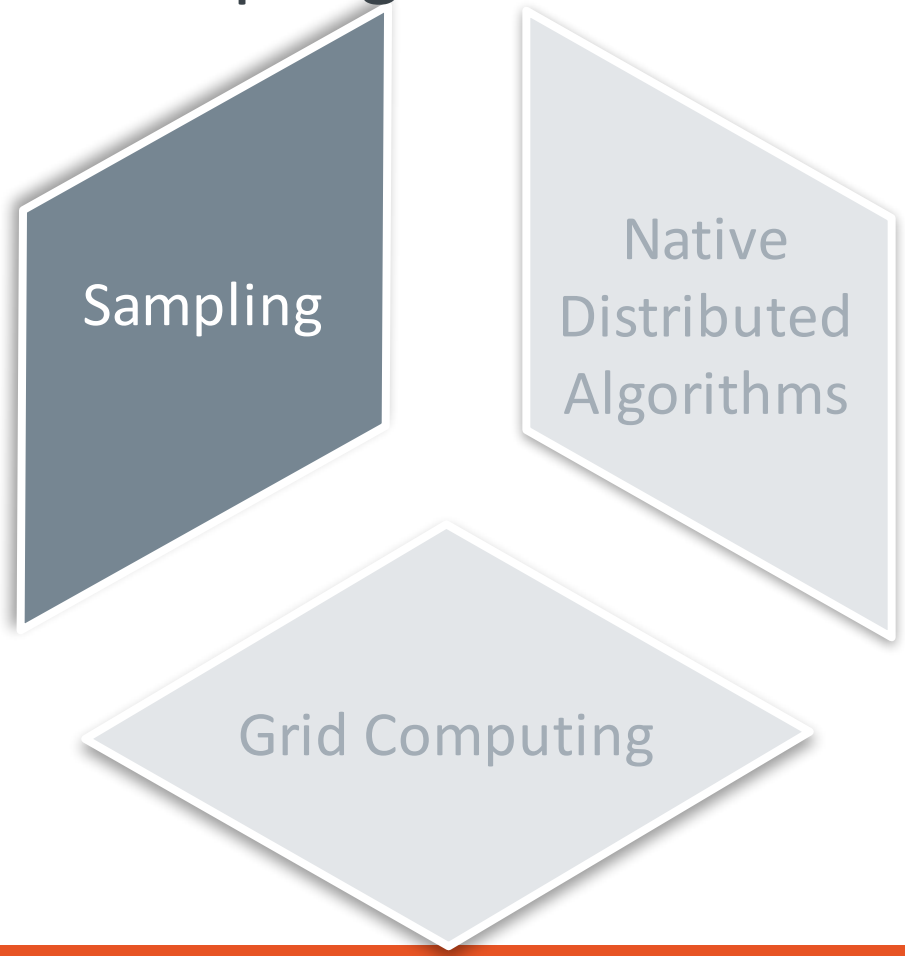


- **SKILLS GAP** is a major adoption inhibitor, cited by 57%
- How to **EXTRACT VALUE** from Hadoop, cited by 49%

Different Approaches to Big Data Analytics

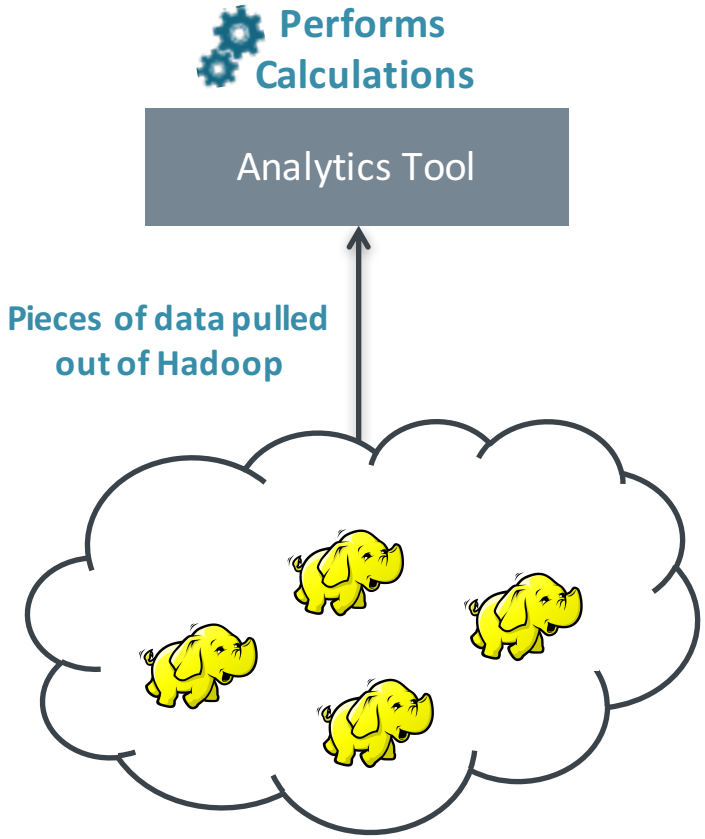


Approach 1: Sampling



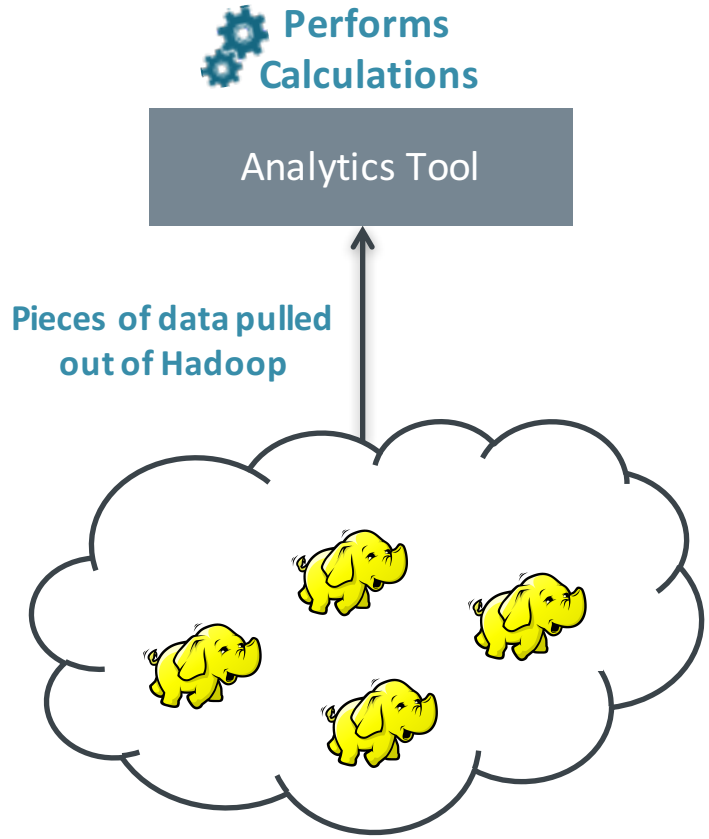
Sampling: Data Movement & Processing

- Data Movement
  - Pulls sample data from HDFS/Hive/Impala
- Data Processing
  - In the analytics tool



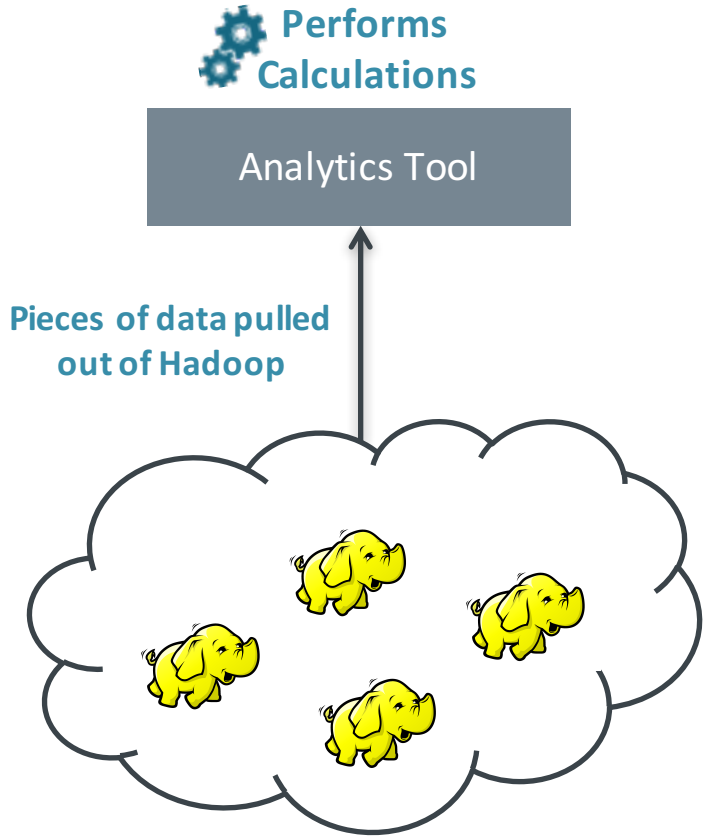
Sampling: Pros & Cons

- Pros
  - + Simple and easy to start with
  - + Usually works well for data exploration and early prototyping
  - + Some ML models would not benefit from more data anyway
- Cons
  - – Many ML models would benefit from more data
  - – Cannot be used when large scale data preparation is needed
  - – Hadoop is used as a data repository only



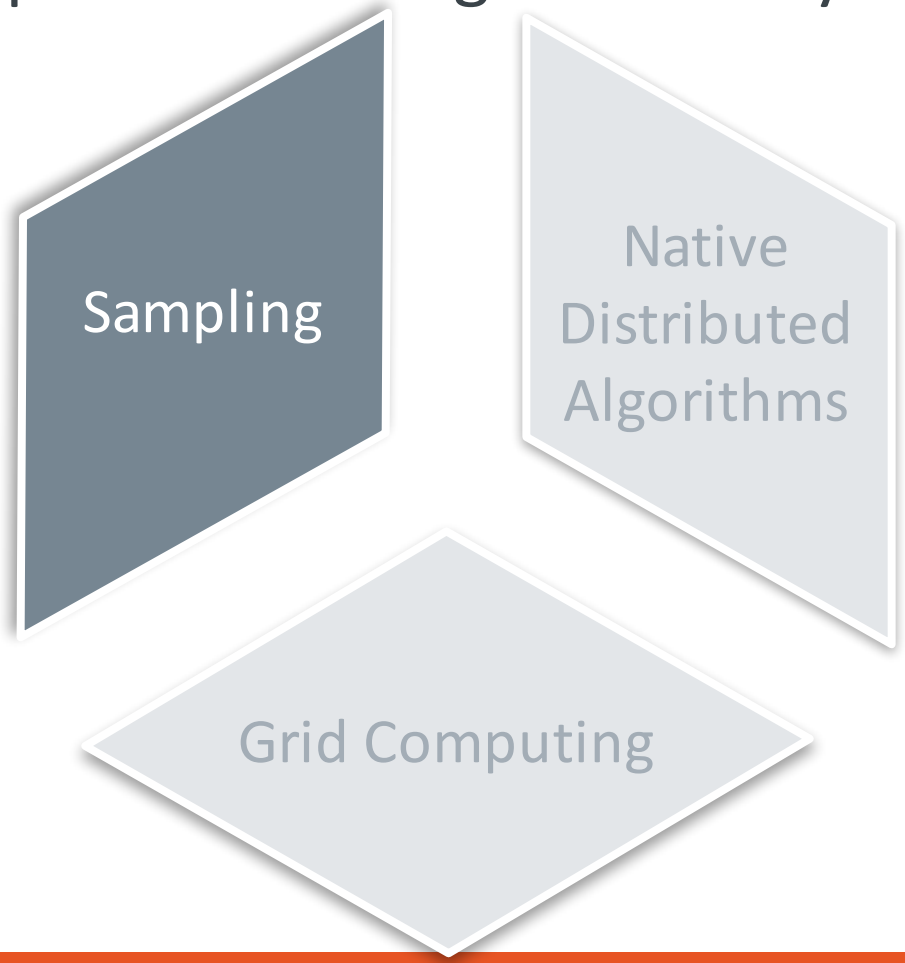
Sampling: Best Practices

- When to use it
  - + Only data exploration / data understanding
  - + Early prototyping on prepared and clean data
  - + Machine Learning modeling with very few and basic patterns (e.g. only a handful of columns and binary prediction target)
- When NOT to use it
  - – Large number of columns in the data
  - – Need to blend large data sets (e.g. large-scale joins)
  - – Complex Machine Learning models
  - – Looking for rare events
- Horror stories
  - – Important decisions made based on biased samples

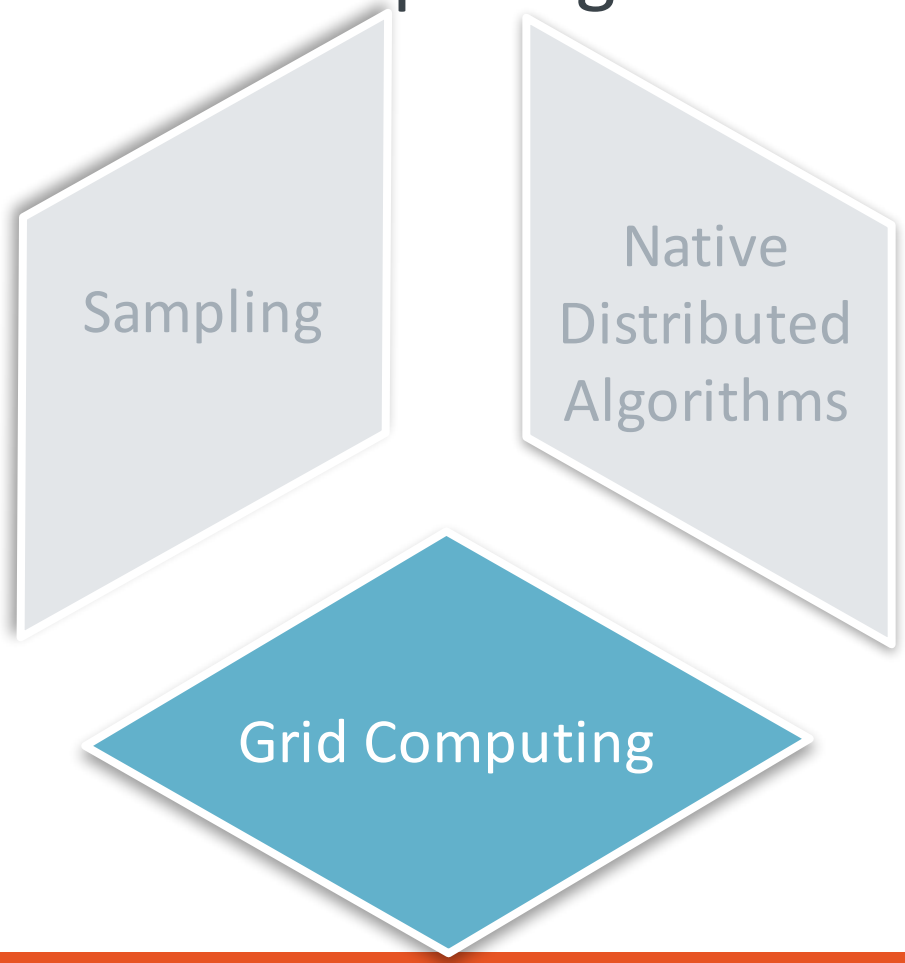


Different Approaches to Big Data Analytics

Data Visualization,
Programming

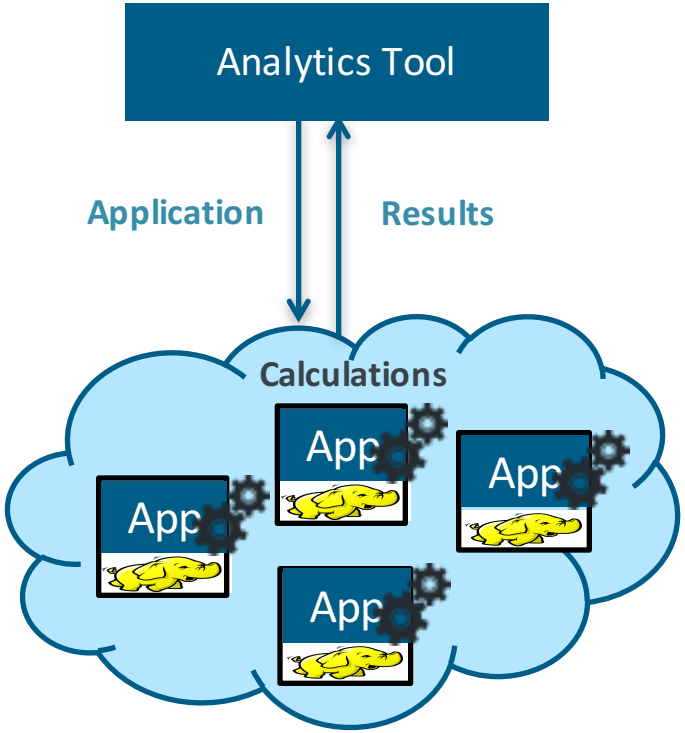


Approach 2: Grid Computing



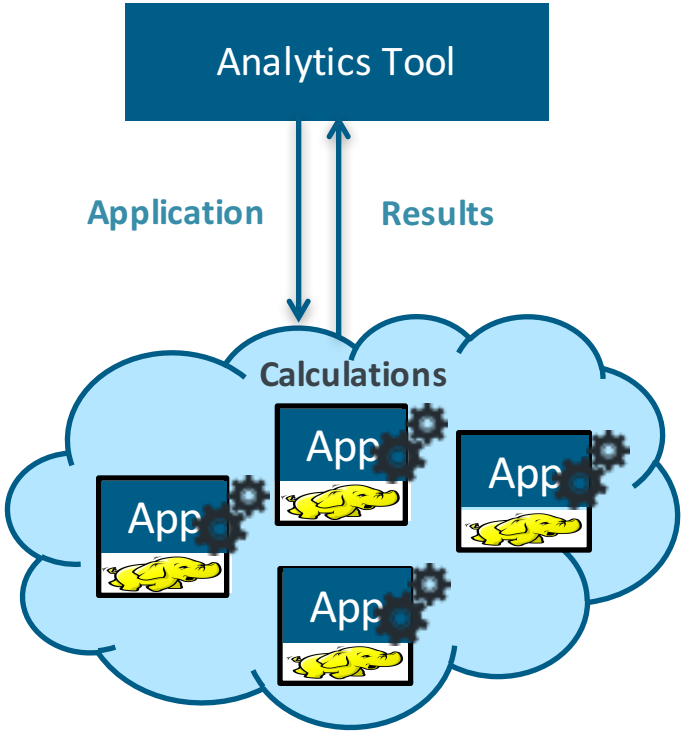
Approach 2: Grid Computing

- *Data Movement*
  - Only results are moved, data remains in Hadoop
- *Data Processing*
  - Custom single-node application running on multiple Hadoop nodes
- *Pros & Cons*
  - \+ Hadoop is used for parallel processing in addition to using as a data source
  - – Only works if data subsets can be processed independently
  - – Only as good as the single-node engine, no benefit from fast-evolving Hadoop innovations

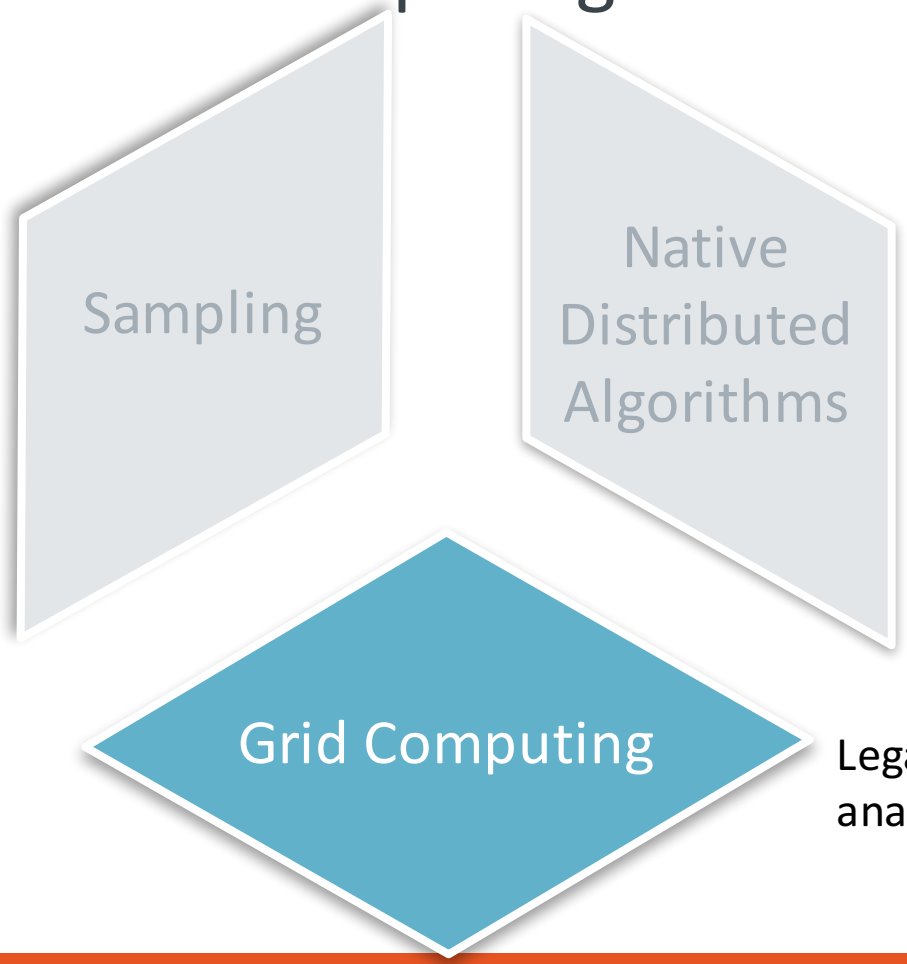


Grid computing best practices

- *When to use it*
  - + Task can be performed on smaller, independent data subsets
  - + Compute-intensive data processing
- *When NOT to use it*
  - – Data-intensive data processing
  - – Complex Machine Learning models
  - – Lots of interdependencies between data subsets
- *Horror stories*
  - – Grid computing job called in huge loops to manage dependencies and intermediate results

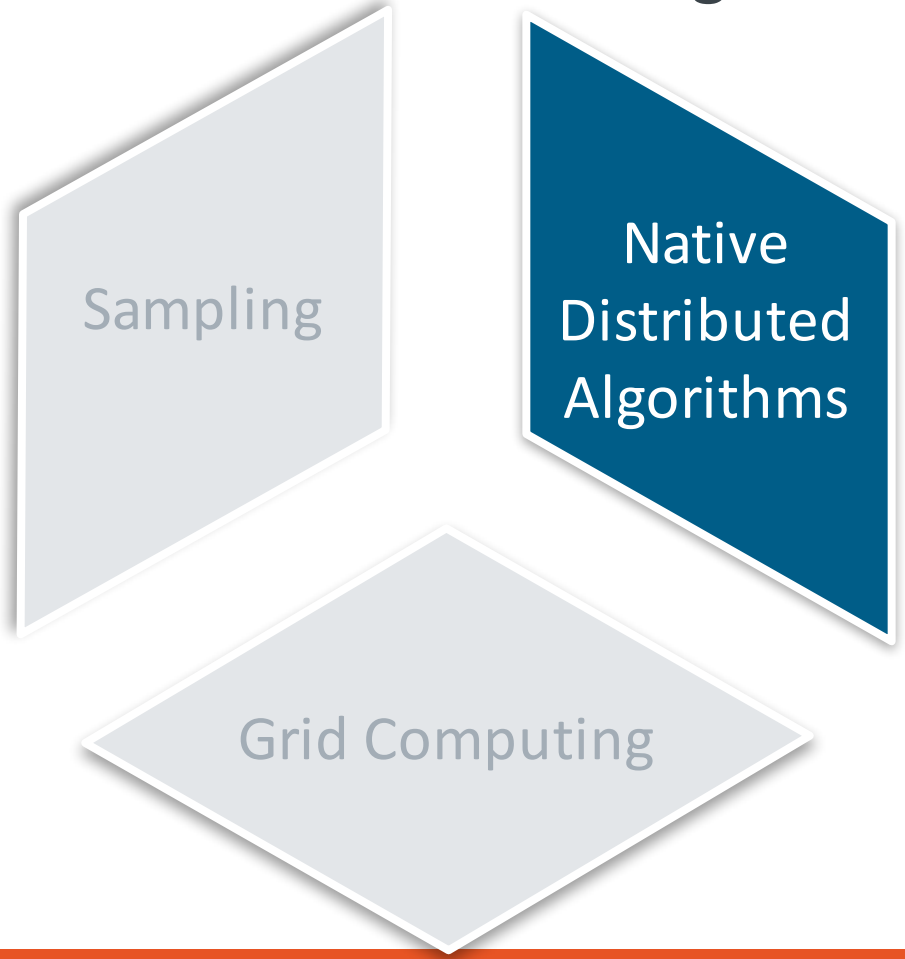


Approach 2: Grid Computing



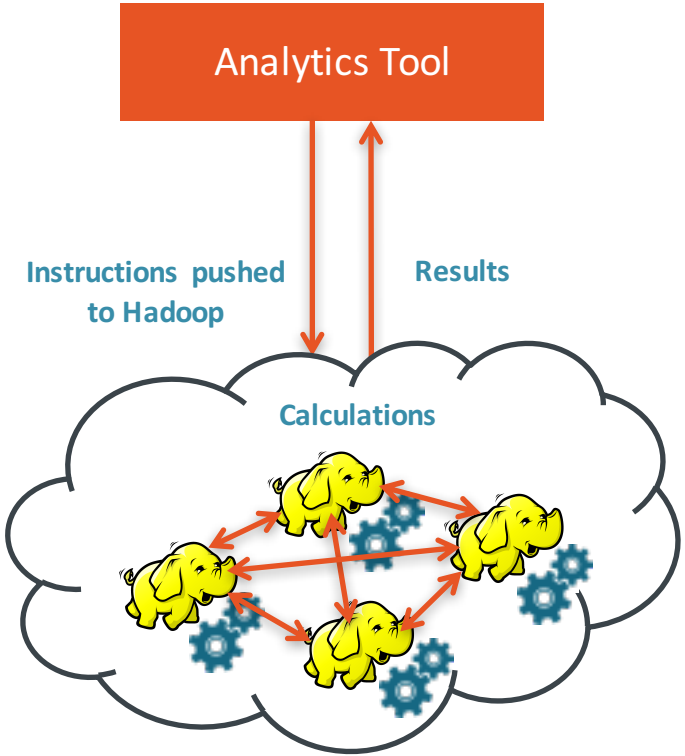
Legacy single-machine analytics engines

Approach 3: Native Distributed Algorithms



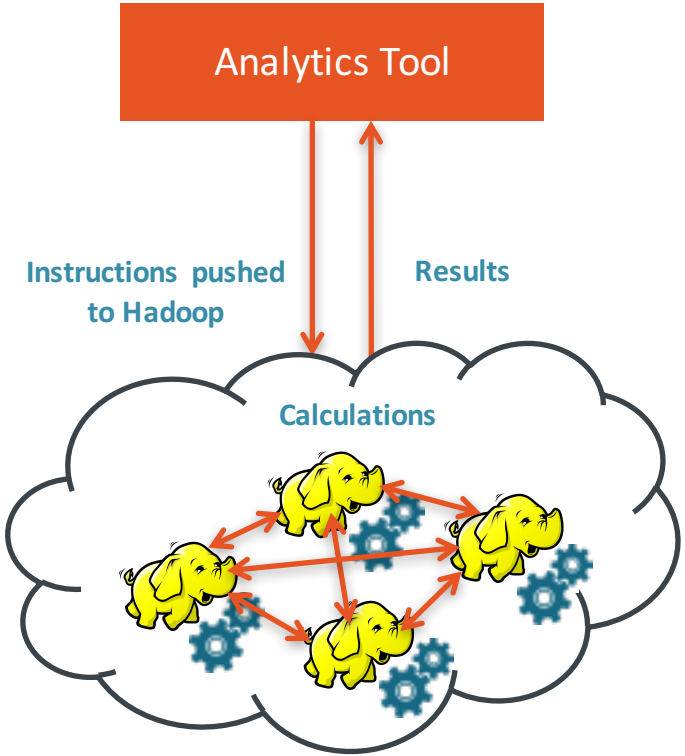
Approach 3: Native distributed algorithms

- *Data Movement*
  - Only results are moved, data remains in Hadoop
- *Data Processing*
  - Executed by native Hadoop tools: Hive, Spark, H2O, Pig, MapReduce, etc.
- *Pros & Cons*
  - \+ Holistic view of all data and patterns
  - \+ Highly scalable distributed processing optimized for Hadoop
  - – Limited set of algorithms available, very hard to develop new algorithms

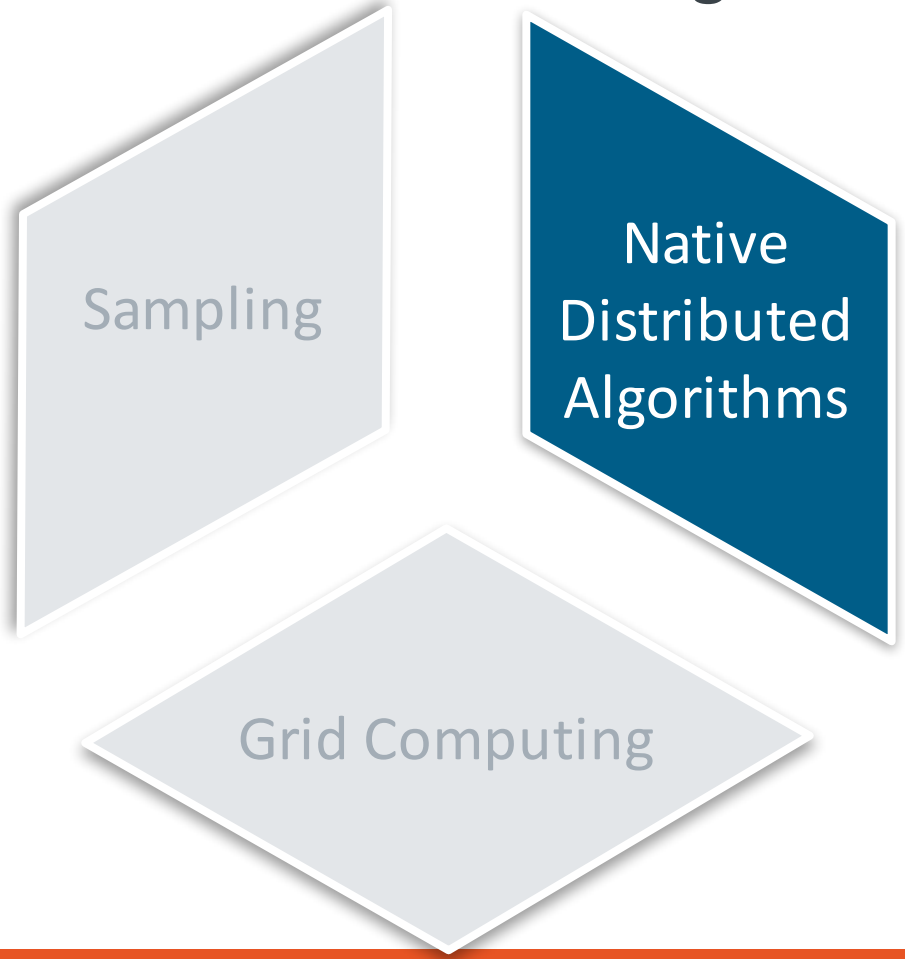


Native distributed algorithms best practices

- *When to use it*
  - + Complex Machine Learning models needed
  - + Lots of interdependencies inside the data (e.g. graph analytics)
  - + Need to blend and cleanse large data sets (e.g. large-scale joins)
- *When NOT to use it*
  - – Data is not that large
  - – Sample would reveal all interesting patterns
- *Horror stories*
  - – Complex ML model developed in 3 months defeated by a prototype model created in an afternoon



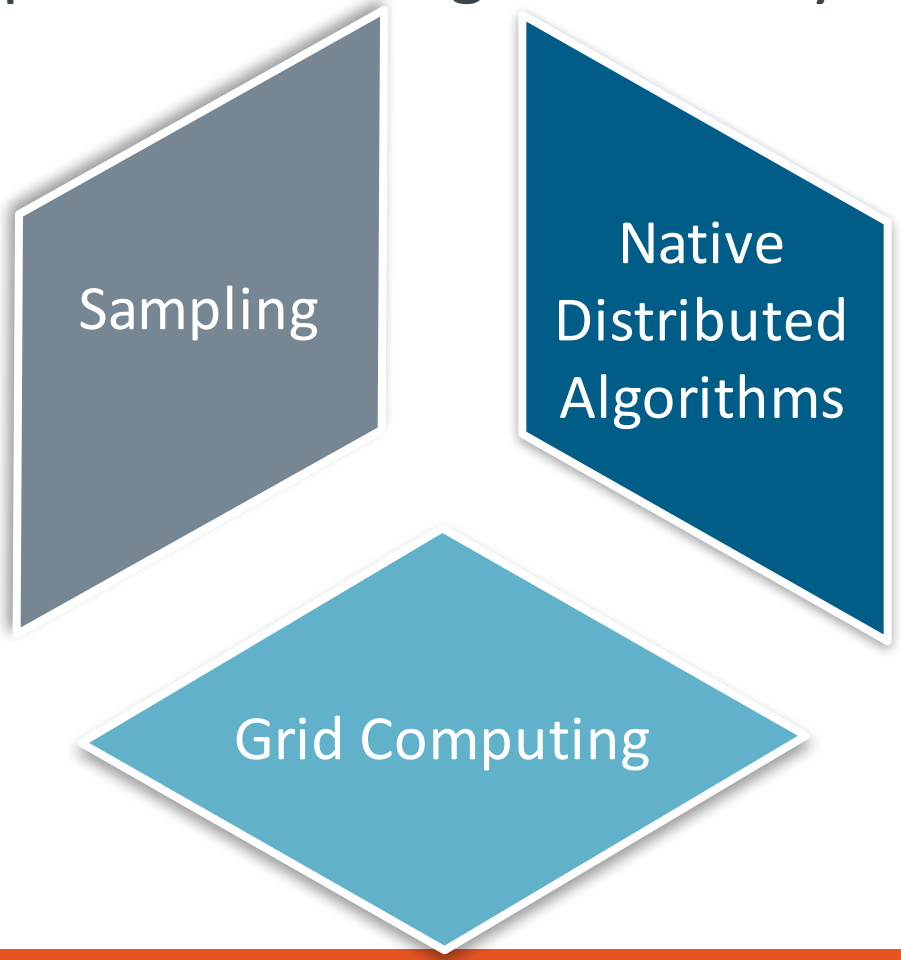
Approach 3: Native Distributed Algorithms



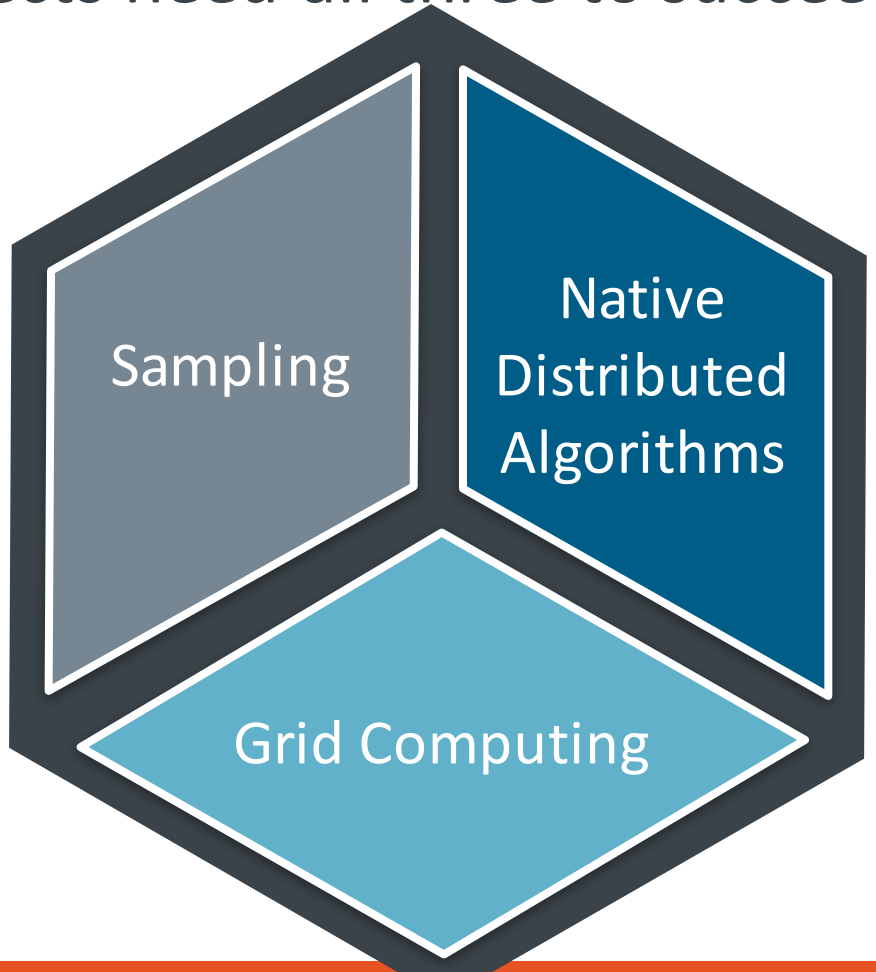
Hadoop ecosystem projects

Different Approaches to Big Data Analytics

Which one to use for a given use case?

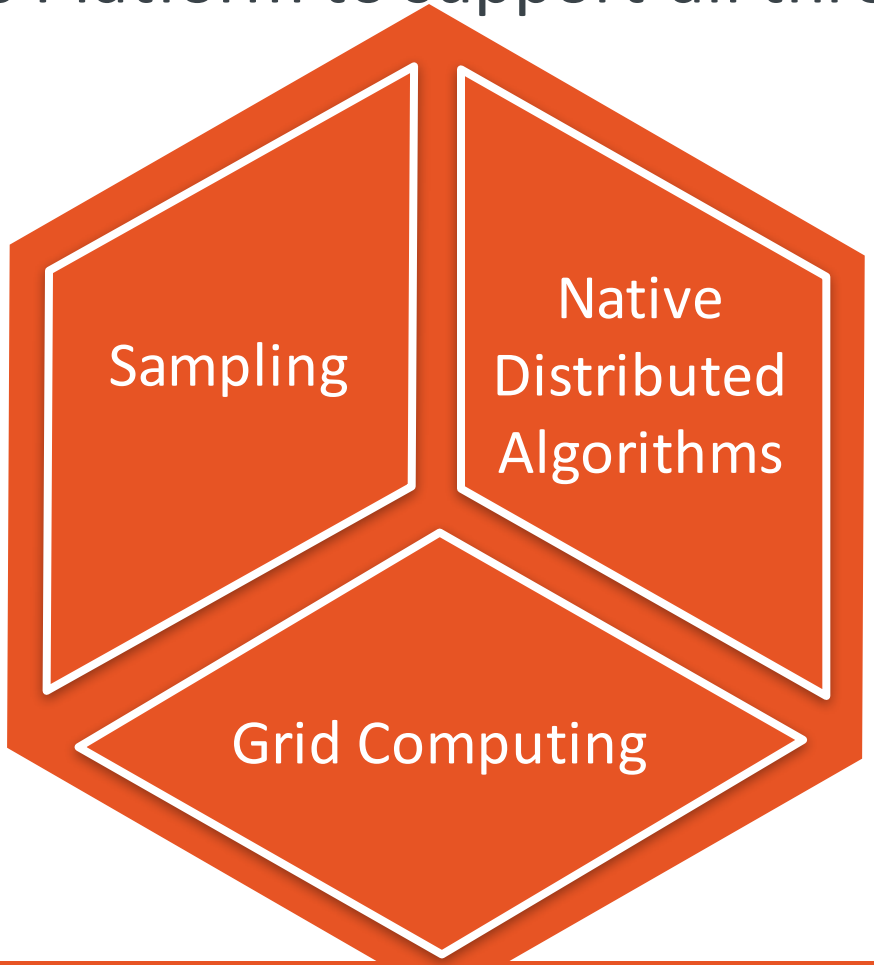


Typical projects need all three to succeed



RapidMiner Predictive Analytics Platform

Single Analytics Platform to support all three



Predictive Analytics *Reimagined*

A Modern Data Science Platform to Turn Data Into a Strategic Asset

Zoltan Prekopcsak

Email: zprekopcsak@rapidminer.com
Twitter: @prekopcsak
Web: rapidminer.com

rapidminer

10 Milk Street, 11th Floor
Boston, MA 02108
USA

+1 617 401 7708