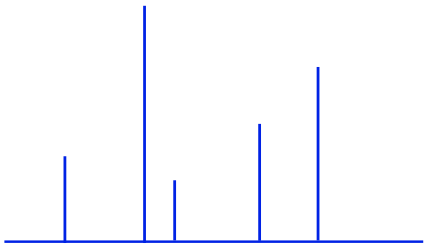
Map/Reduce Uncollapsed Gibbs Sampling for Bayesian Nonparametric Models

Melanie F. Pradier, Pablo G. Moreno, Francisco J. R. Ruiz,
Isabel Valera, Harold Molina-Bulla and Fernando Perez-Cruz

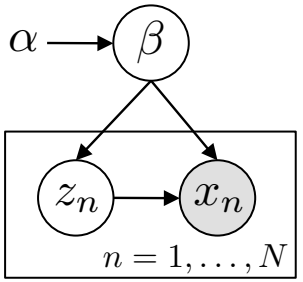
- BNP models were developed for *big data*.
- . . . but inference is still slow.
- We need to scale up inference for BNP to be applicable.

Our approach

- Our approach: Instantiate the latent measure.
  - Atom weights.
  - Atom locations.

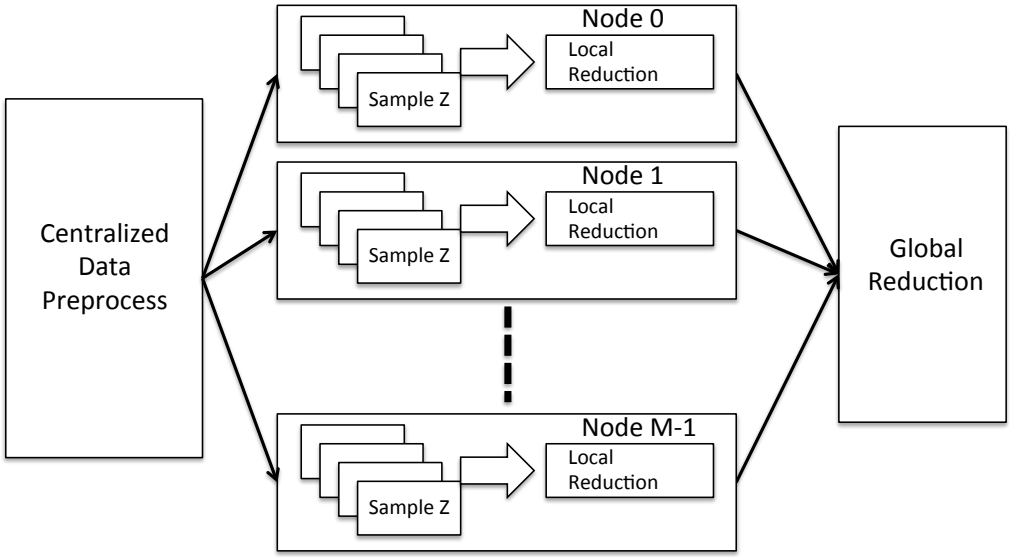


- Conditioned on it, local variables are independent.

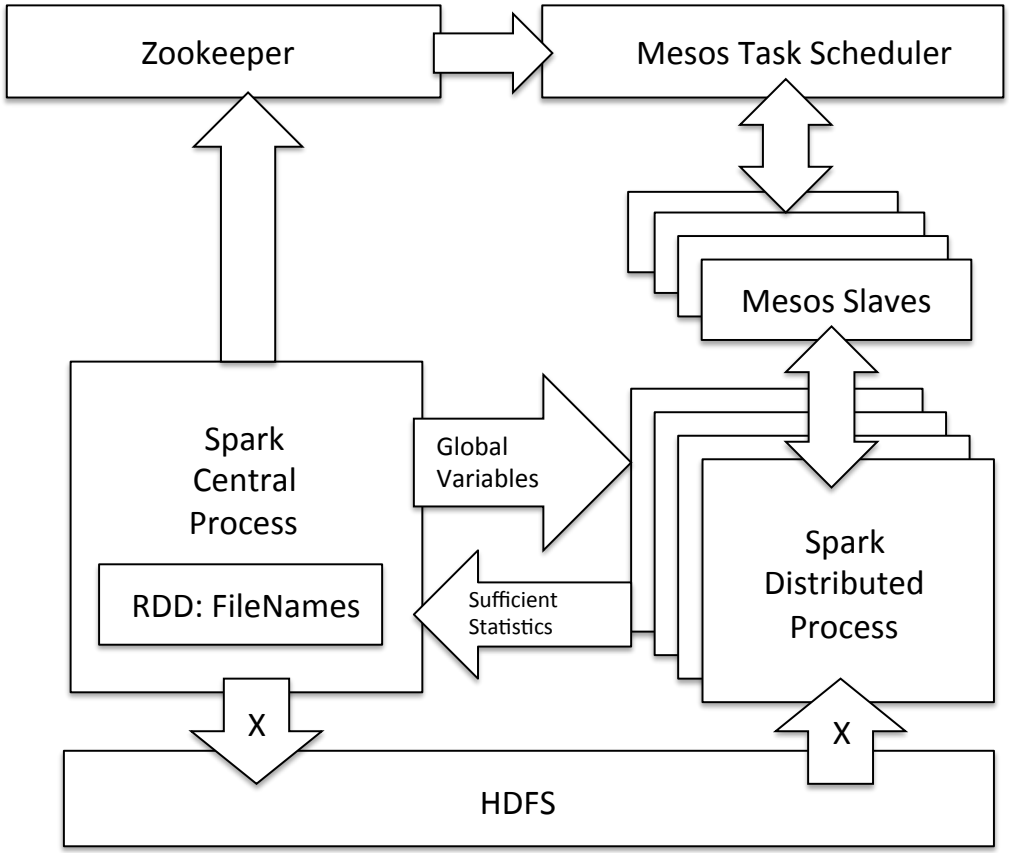


Our approach

- Generic Scala code.
- Map/Reduce scheme.
- Two implementations:
  - Parallel implementation.
  - Distributed implementation.
    - Apache Spark.
    - Hadoop file system.

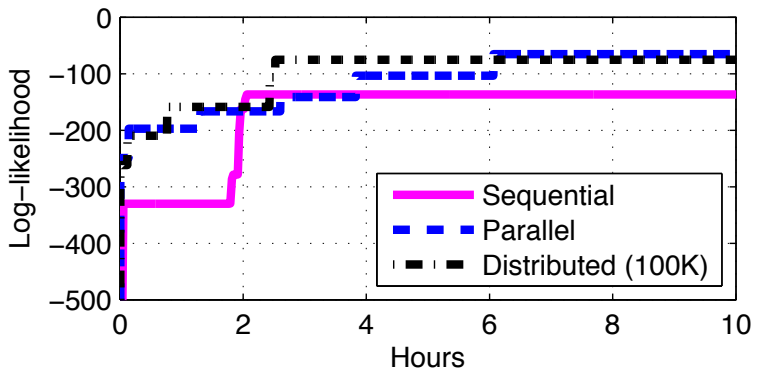


Architecture



Experiments

Synthetic dataset (dimensionality $D = 10$)
1M observations (up to 50M)



Algorithm \ N	100K	1M	5M	50M
Sequential	0.1349	1.3963	-	-
Parallel	0.0123	0.1397	0.8736	-
Distributed (100K)	0.1795	0.1512	0.2143	**1.3429**

Time (minutes) per iteration

Thank you for you attention!