MAP/REDUCE UNCOLLAPSED GIBBS SAMPLING FOR BAYESIAN NON PARAMETRIC MODELS

MELANIE F. PRADIER[1], PABLO G. MORENO[1], FRANCISCO J. R. RUIZ[1],
ISABEL VALERA[1], HAROLD MOLINA-BULLA[1] AND FERNANDO PEREZ-CRUZ[1,2]
1. University Carlos III in Madrid, Spain and 2. Bell Labs, Alcatel-Lucent, New Jersey, USA
{melanie,pgmoreno,franrruiz,ivalera,hmolina,fernando}@tsc.uc3m.es

INTRODUCTION

Bayesian non-parametric (BNP) models are desirable tools for *big data* analysis, not only because of accuracy and flexibility, but because they can provide *interpretable* explanations of the data. Yet, inference is typically hard to scale.

Observations

- Two general trends: Variational Inference or MCMC Sampling
- Either exact or approximate approaches
- State-of-the-art methods = quite similar, often task-specific

This paper presents:

- A general framework for parallel MCMC inference in BNP models.
- Modular code in Spark/Scala, easily extensible to other likelihood functions.
- Particular implementations for the Dirichlet and Beta Processes.

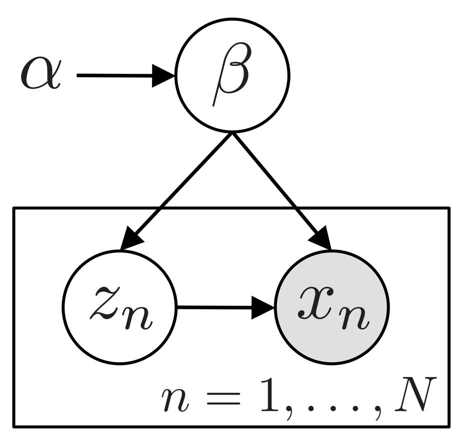
GENERAL APPROACH

The joint distribution of the observations and latent variables is given by

$$p(\beta, \mathbf{x}, \mathbf{z}|\alpha) = p(\beta|\alpha) \prod_{n=1}^{N} p(x_n, z_n|\beta), \tag{1}$$

where $\mathbf{x} = \{x_1, \dots, x_N\}$ and $\mathbf{z} = \{z_1, \dots, z_N\}$.

According to *De Finetti's theorem*, if observations $x_1, \dots, x_N$ are exchangeable, there exists a **latent measure** that makes such observations conditionally independent.



1. Sample β from $p(\beta|\mathbf{z}, \mathbf{x}, \alpha)$.
2. Sample z_n from $p(z_n|x_n, \beta)$ for $n = 1, \dots, N$.

INFERENCE FOR DIRICHLET PROCESS

- Samplers for both DP and BP process based on stick-breaking representation.
- At each iteration, propose T new clusters with parameters drawn from the prior. The weights assigned to each cluster are given by

$$\pi \sim \text{Dirichlet}\Big(N_1, \dots, N_{K_+}, \underbrace{\frac{\gamma}{T}, \dots, \frac{\gamma}{T}}_{T \text{ times}}\Big), \tag{2}$$

INFERENCE FOR BETA PROCESS

1. Sample total mass for all the sticks

$$S = \sum_{k=1}^{\infty} \pi_k. \tag{3}$$

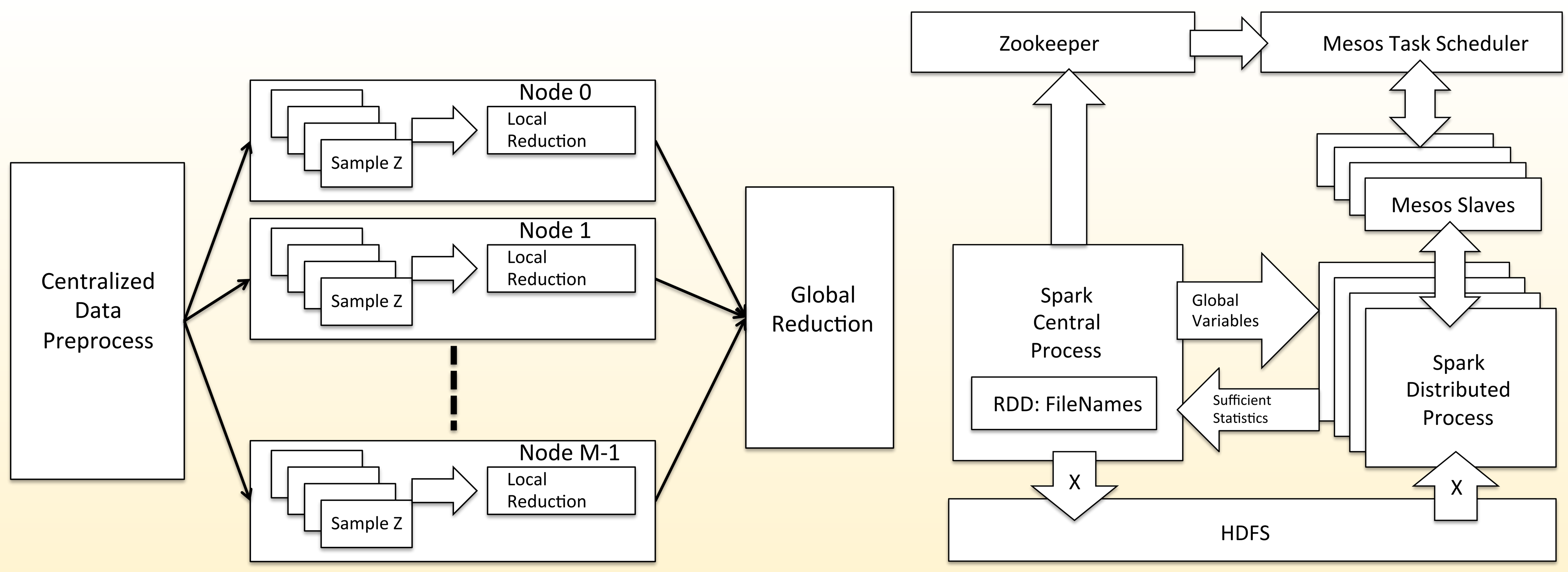
2. Distribute remaining probability mass among new sticks. We follow an approximate random procedure fulfilling two constraints:
   - sum of all sticks can never surpass S
   - new sticks must be smaller than "already active" sticks

ALGORITHM

Algorithm 1 Distributed parallel sampling.

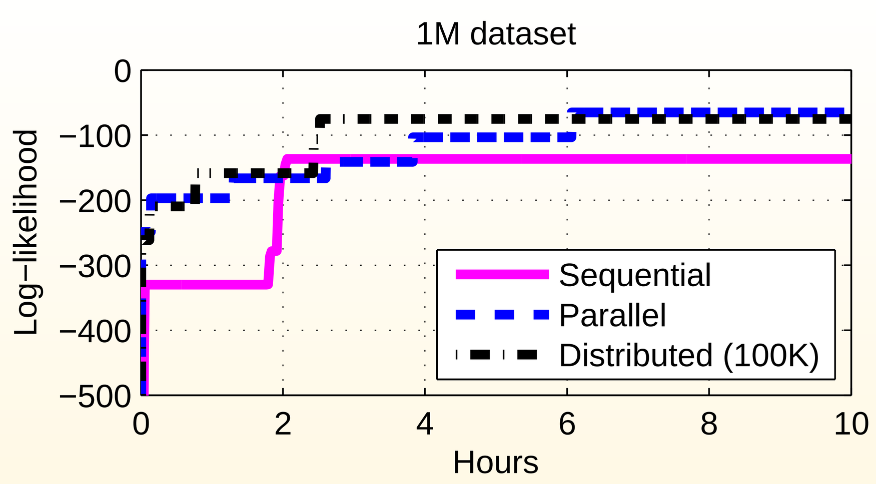
Split the data into chunks.
Store the pieces in HDFS.
Initialize the task scheduler.
while it<maxIter **do**
  Sample the shared variable β.
  Distribute to each node: chunk reference and β.
  Sample the per-datum variables in each node and return sum of local sufficient statistics.
  Apply reduce operator to join all the sub-results.
  Clean empty clusters/features.
end while

ARCHITECTURE



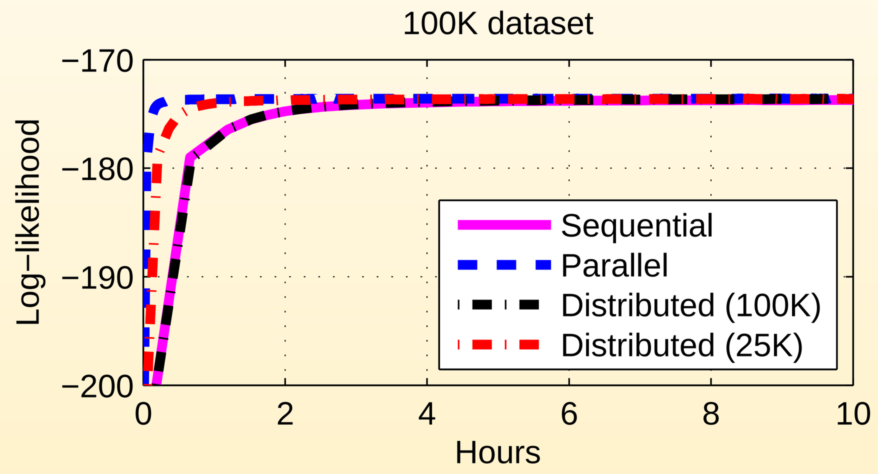
RESULTS FOR DIRICHLET PROCESS

Synthetic data:



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

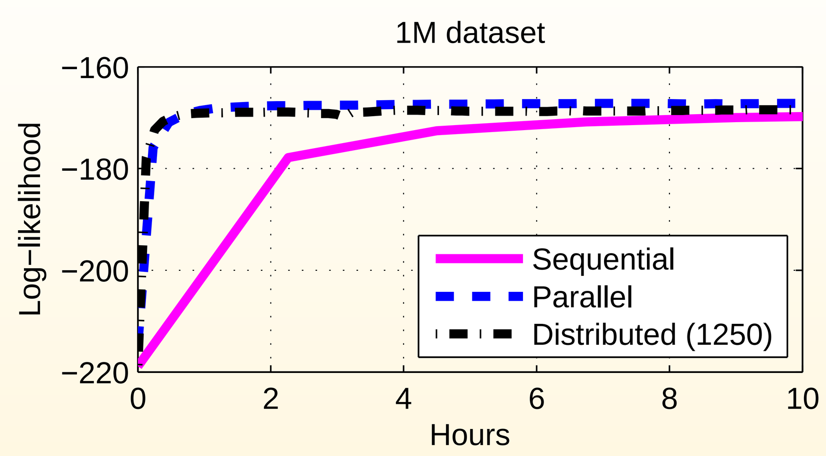
Real data (tinyImages):



Algorithm \ N	100K	1M
Sequential	9.9169	-
Parallel	0.6791	26.4195
Distributed (100K)	10.7375	32.8588
Distributed (25K)	3.1215	9.6388

RESULTS FOR BETA PROCESS

Synthetic data:



Algorithm \ N	100K
Sequential	122.86
Parallel	9.73
Distributed (1,250)	5.17

CONCLUSIONS AND FUTURE WORKS

Contributions

- Powerful software to parallelize inference in BNP models.
- General framework and flexibility code.
- Parallel and distributed implementations for DP and BP.

Future Work

- Extensive empirical analysis of sampler properties
- More efficient data-driven proposals
- Parallelization for collapsed Gibbs Sampling

FUNDING

- *This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861*