UNIVERSITÉ LIBRE DE BRUXELLES
Faculté des Sciences
Département d'Informatique

INFO-H-515
Big Data Scalable Analytics

Report - Phase 2

Antoine Passemiers

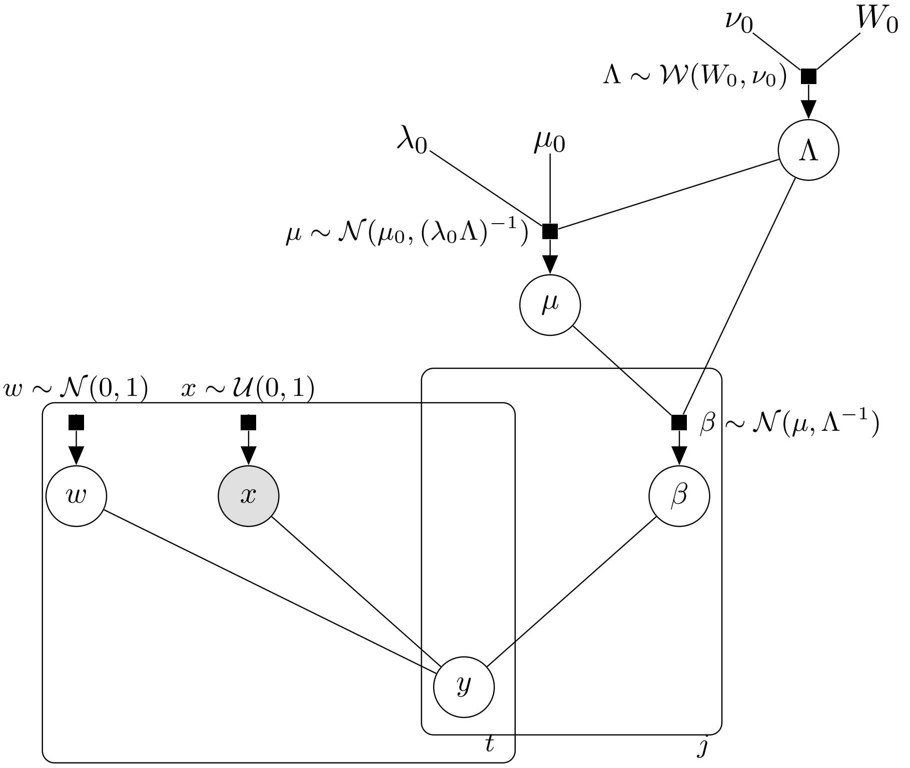
Academic year 2018 - 2019

CONTENTS

1. INTRODUCTION

2. ARCHITECTURE

2.1 Data generation


ν_0
W_0
$\Lambda \sim \mathcal{W}(W_0, \nu_0)$
Λ
λ_0
μ_0
$\mu \sim \mathcal{N}(\mu_0, (\lambda_0\Lambda)^{-1})$
μ
$w \sim \mathcal{N}(0,1)$
$x \sim \mathcal{U}(0,1)$
$\beta \sim \mathcal{N}(\mu, \Lambda^{-1})$
w
x
β
y
t
j

2.2 Recursive Least Squares (RLS) with forgetting factor

In the standard RLS implementation with forgetting factor, the weights β are estimated incrementally using the following formulas:

$$\begin{cases} V^{(t)} & = \frac{1}{\nu}\left(V^{(t-1)} - \frac{V^{(t-1)}x_t^T x_t V^{(t-1)}}{1+x_t V^{(t-1)}(x_t^T} \right) \\ \alpha^{(t)} & = V^{(t)}x_t^T \\ e & = y^{(t)} - x_t\hat{\beta}^{(t-1)} \\ \hat{\beta}^{(t)} & = \hat{\beta}^{(t-1)} + \alpha^{(t)}e \end{cases} \tag{2.1}$$

where

First approach – Fully-vectorized version:

$$\begin{cases} \alpha_t & = V^{(t)} x_t^T \\ e & = y^{(t)} - x_t \hat{B}^{(t-1)} \\ \hat{B}^{(t)} & = \hat{B}^{(t-1)} + \alpha_t^T e \end{cases} \tag{2.2}$$

Second approach – Distributed version:

$$\begin{cases} \alpha_t & = V^{(t)} x_t^T \\ e & = y_j^{(t)} - x_t \hat{B}_{\cdot j}^{(t-1)} \\ \hat{B}_{\cdot j}^{(t)} & = \hat{B}_{\cdot j}^{(t-1)} + \alpha_t^T e \end{cases} \tag{2.3}$$

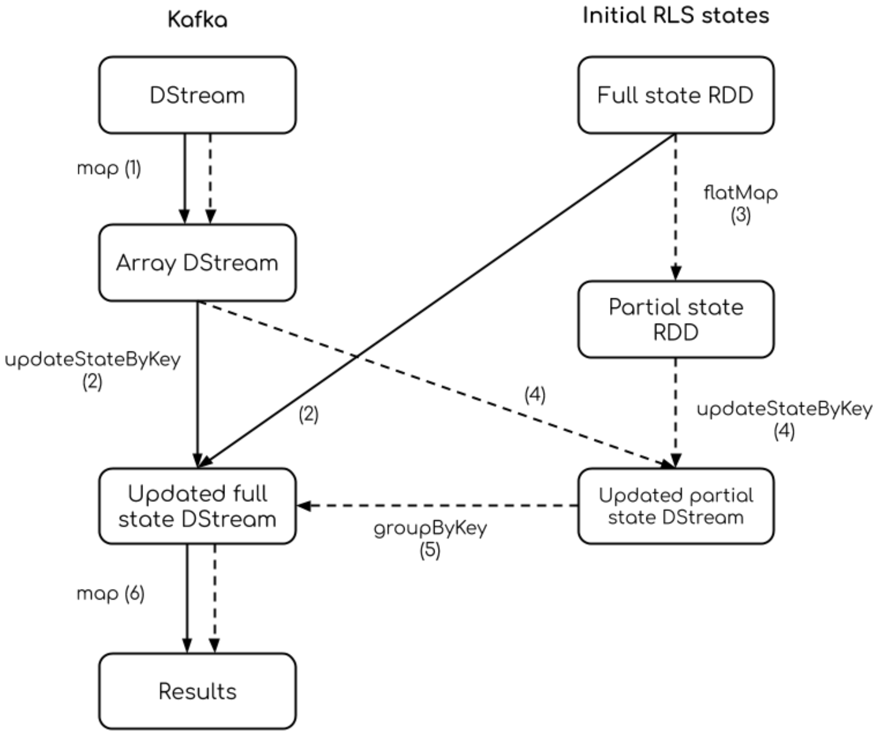


Fig. 2.1: Lineage graph for the proposed architecture. Dashed lines indicate the Spark transformations applied in the distributed version and plain lines indicate the transformations applied in the fully-vectorized version.

3. SCALABILITY

Scalability is the ability of a big data system to process an increasing amount of incoming data by having recourse to an increasing amount of resources. It can be measured by the scaleup, the ratio between the amount of data processed by the model with two different amount of resources but while being run for the same amount of time.

TODO: sub-linear scaleup?