KMeans++

Group 06

Mingliang Qi	360303
Mingyuan Wu	352043
Yuwen Chen	352038

Fachgebiet Datenbanksysteme und Informationsmanagement
Technische Universität Berlin
http://www.dima.tu-berlin.de/

Outline

- K-Means
  - Algorithm
  - Problem: Poor Initial Centroids
- K-Means++
  - Algorithm
  - Stratosphere implementation
  - Spark implementation
  - Experiments and Analysis

K-Means Algorithm

Data: a set of observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$,
number of clusters k, Convergence Delta ξ
Result: a set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

Initialization:

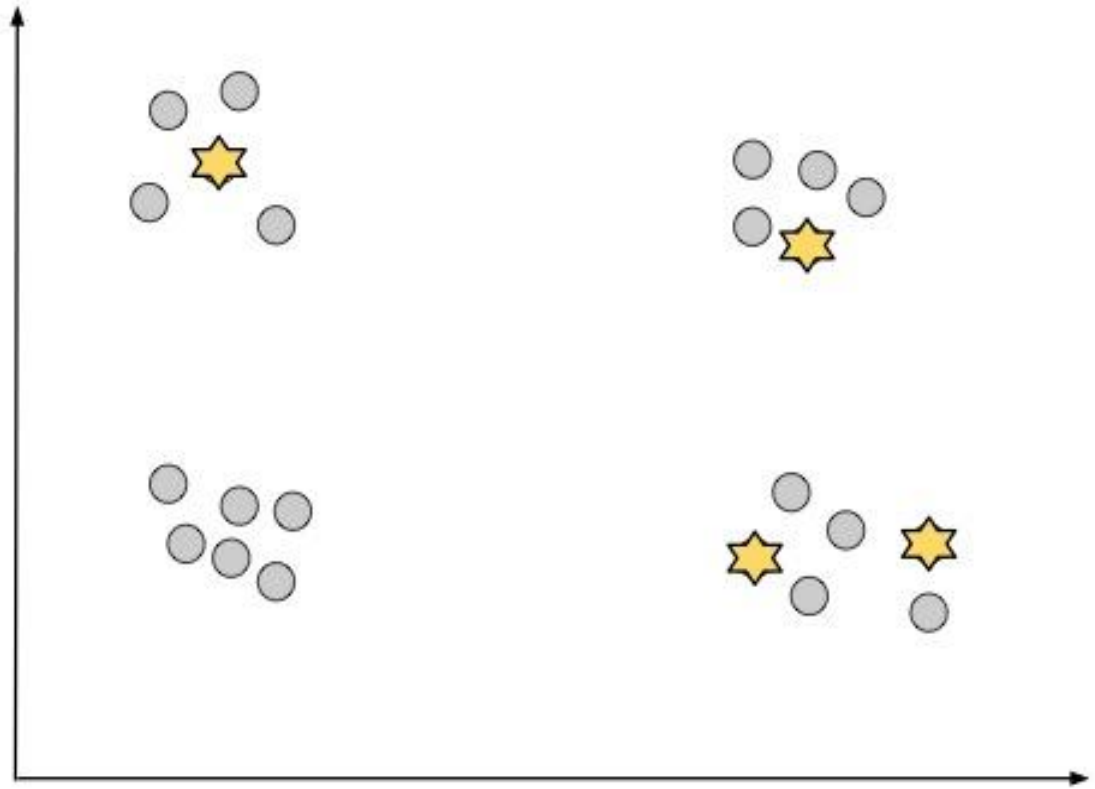
select uniformly k data points $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n\}$ as the centroids of clusters

Compute:

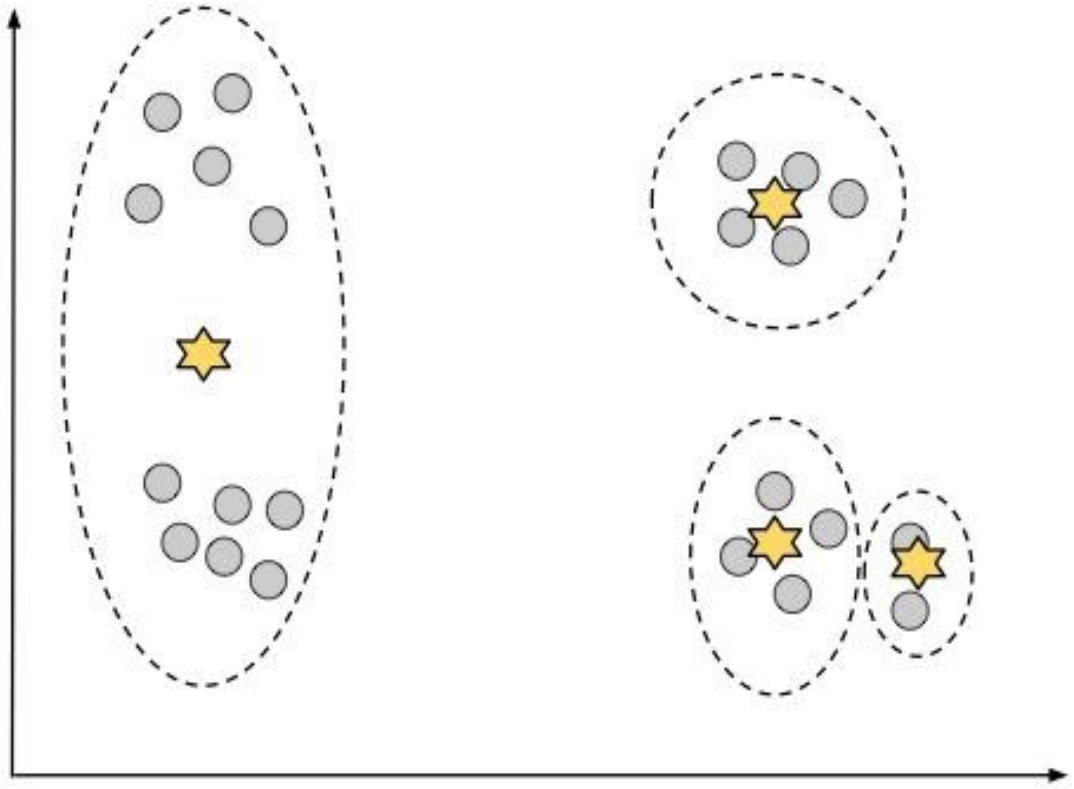
repeat
Form k clusters by assigning each point to its closest centroid;

Recompute the center of each cluster;
until $\Delta\mathcal{C} < \xi$;

Problem: Poor Initial Centroids



Problem: Poor Initial Centroids



KMeans++ Initialization

Data: a set of observations $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$,
number of clusters k, Convergence Delta ξ

Result: a set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

Initialization:

Take a centroid $\mathbf{c}_1$, chosen uniformly at random from $\mathcal{X}$

repeat

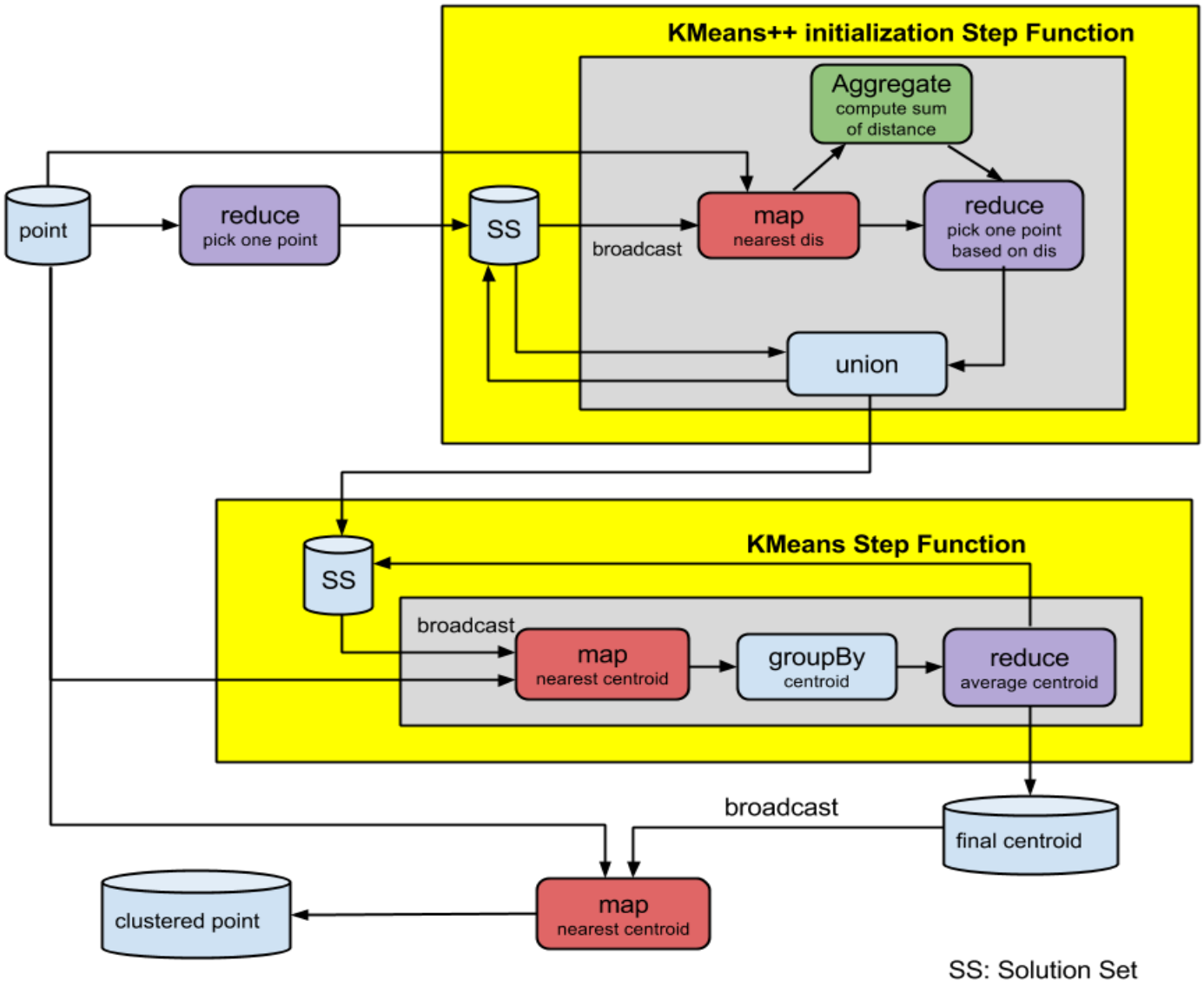
Take a new center $\mathbf{c}_i$, choosing $\mathbf{x} \in \mathcal{X}$ with probability $\frac{D(\mathbf{x})^2}{\sum_{\mathbf{x} \in \mathcal{X}} D(\mathbf{x})^2}$

until *k centroids generated* ;

Compute:

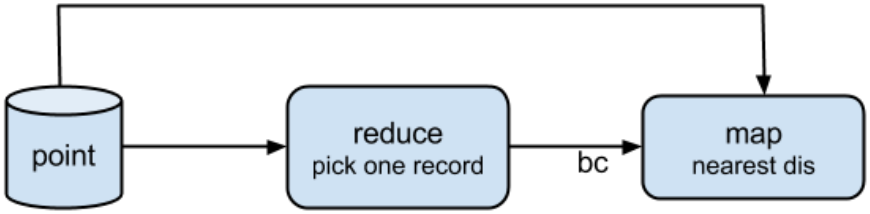
Proceed as with the standard k-means algorithm

Stratosphere implementation

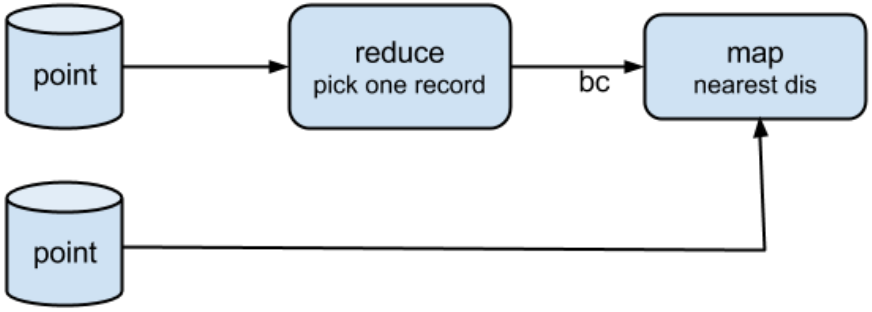


Problem in Stratosphere

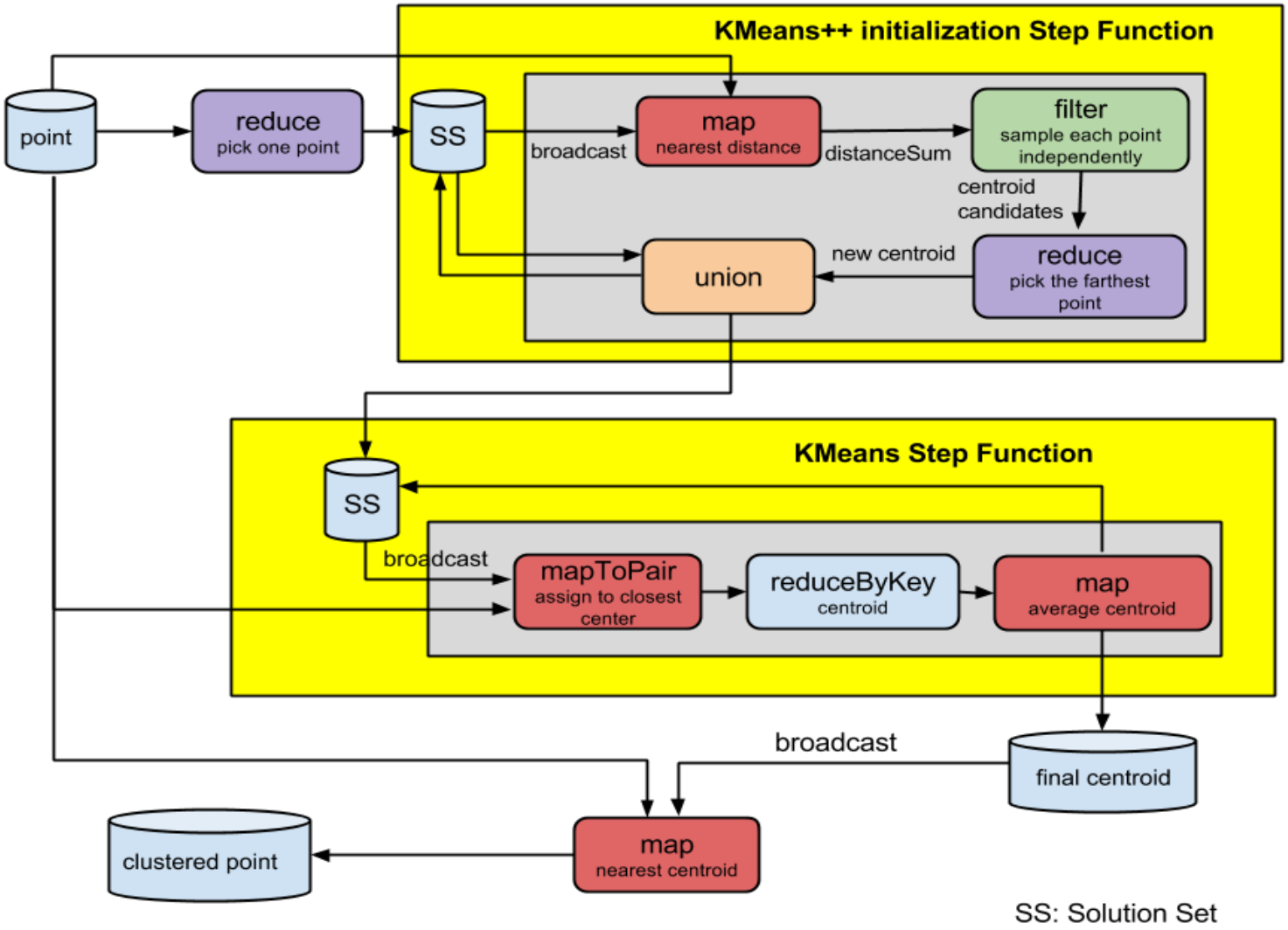
Broadcast, cross operator doesn't work properly sometimes



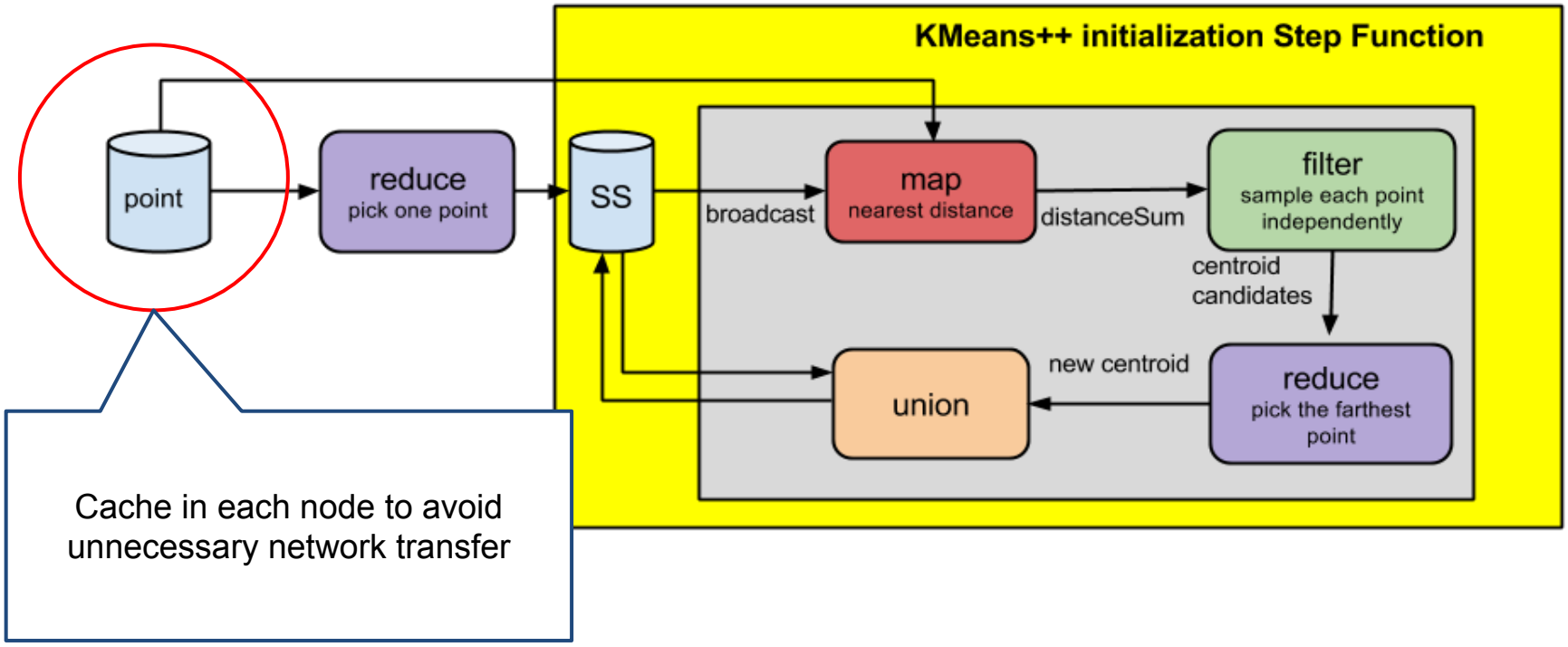
Duplicate the data source works



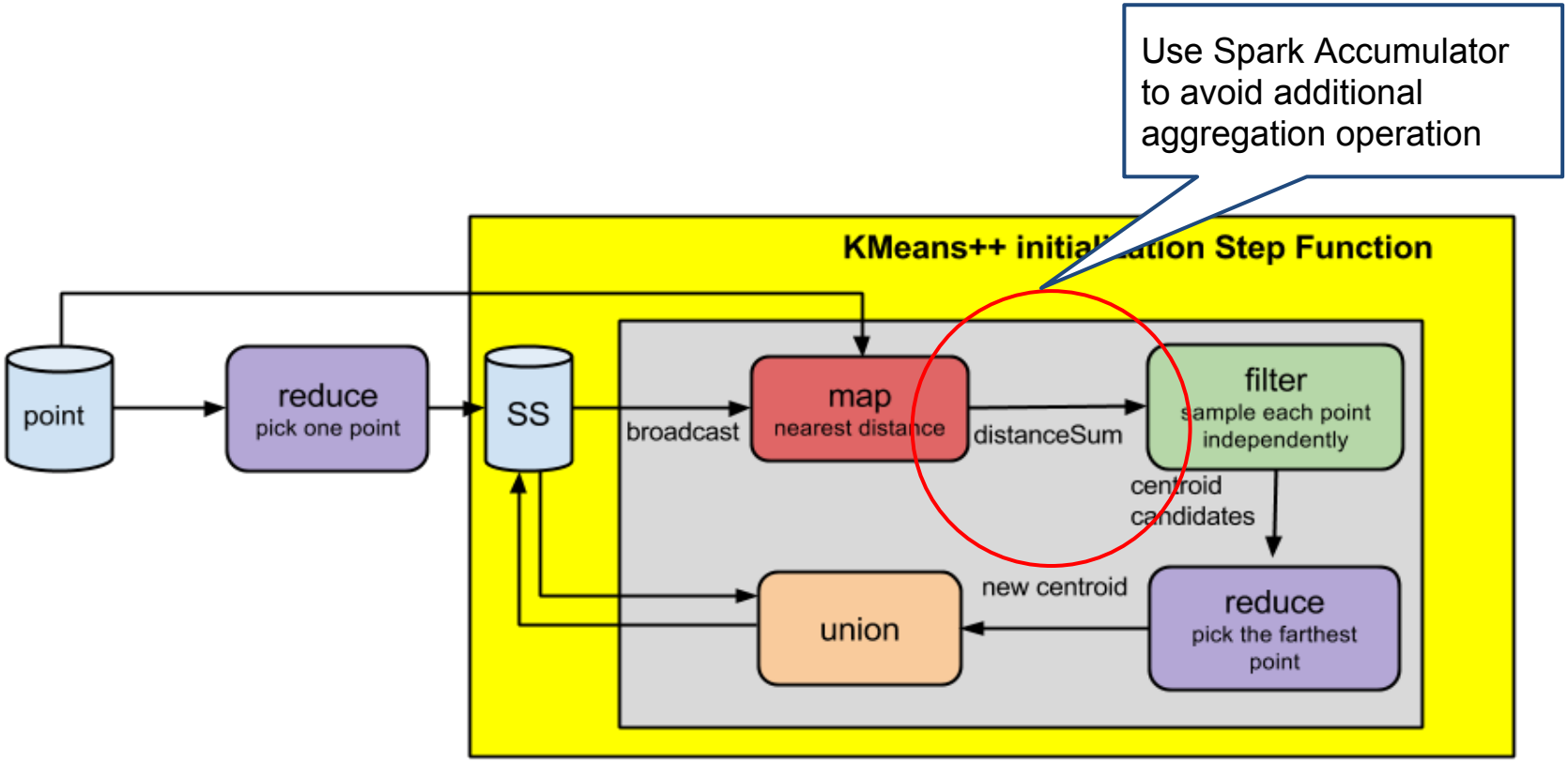
Spark Implementation



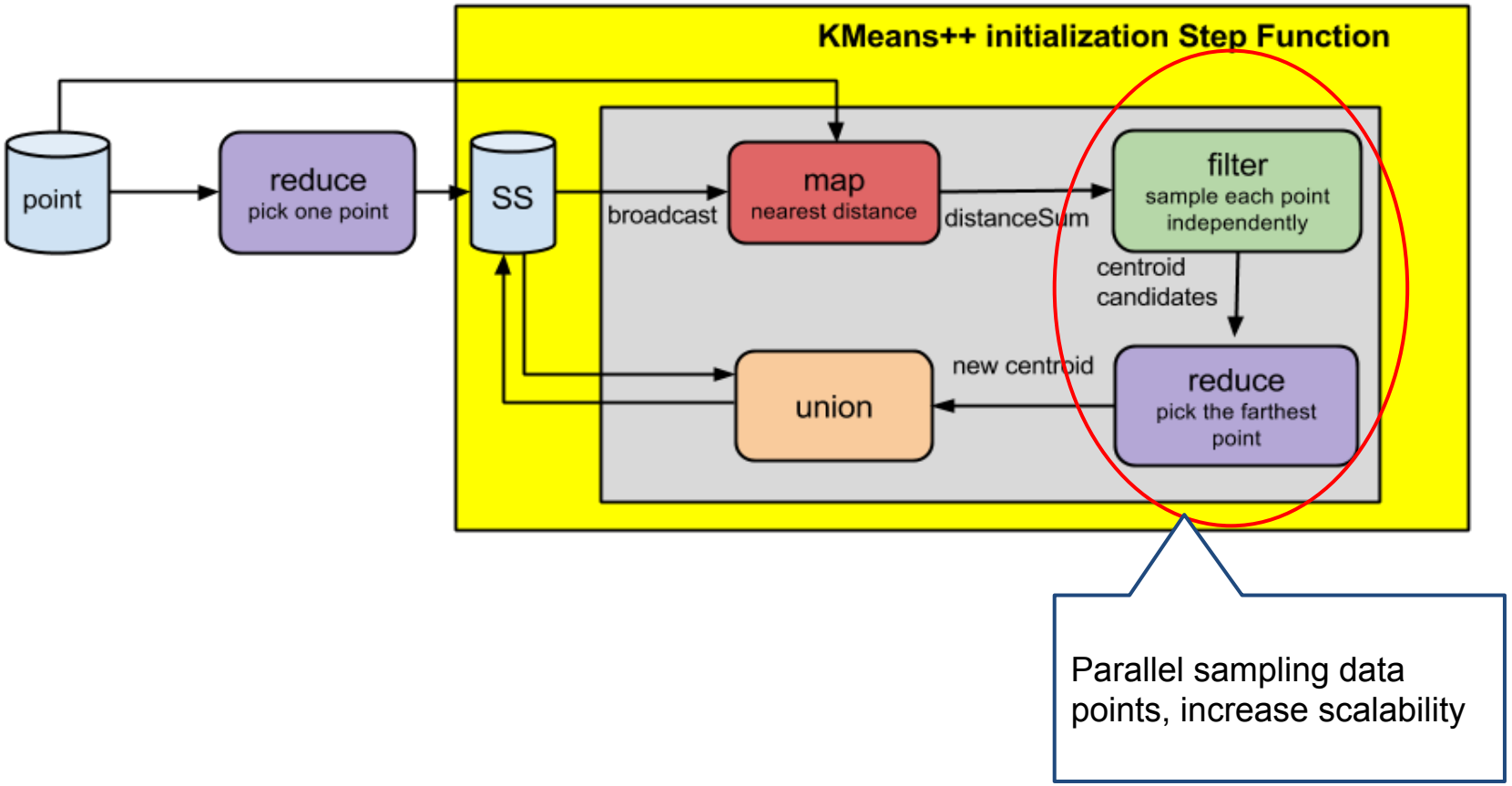
Spark Implementation



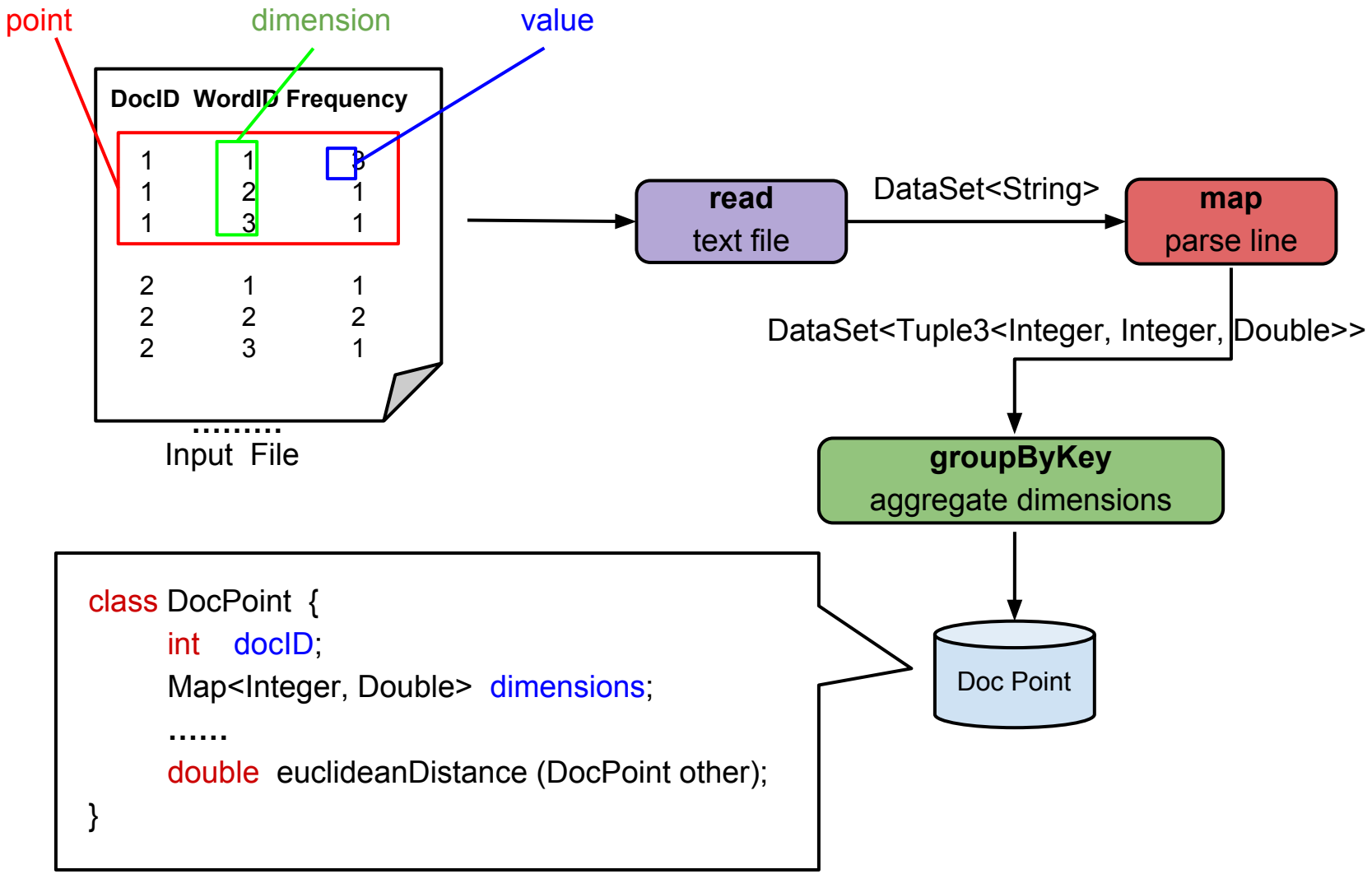
Spark Implementation



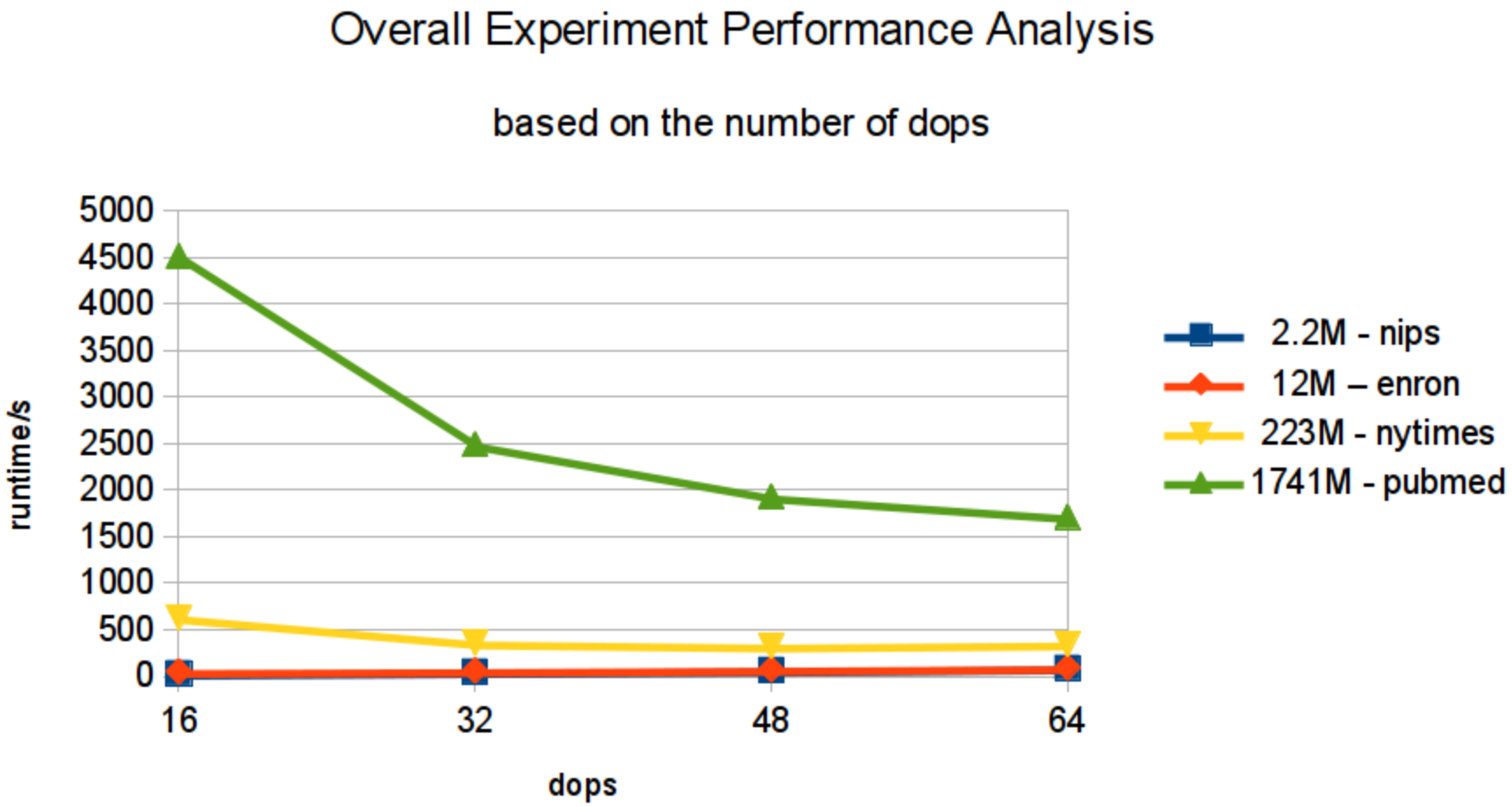
Spark Implementation



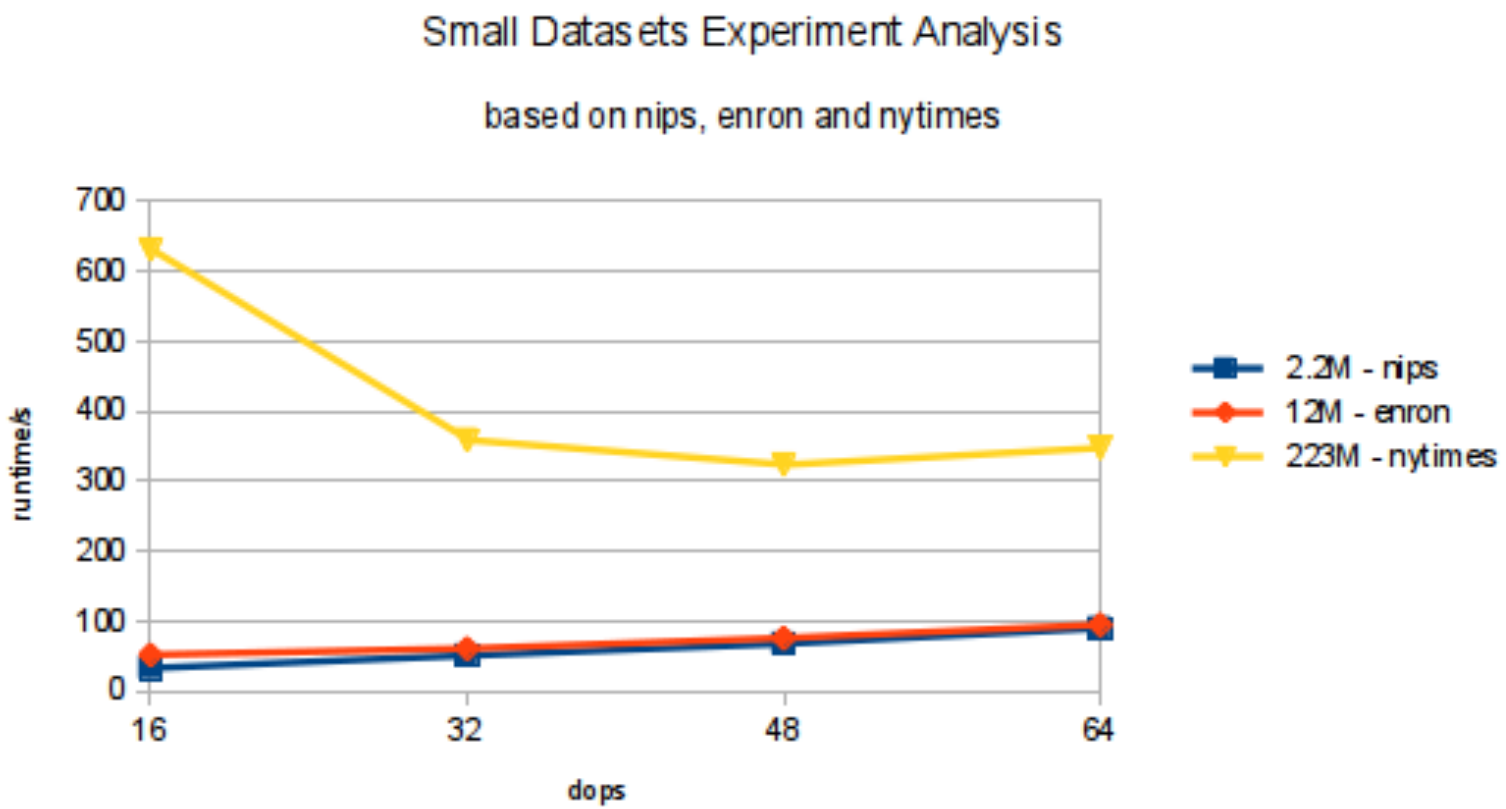
Experiments and Analysis - Bag Of Words



Experiments and Analysis



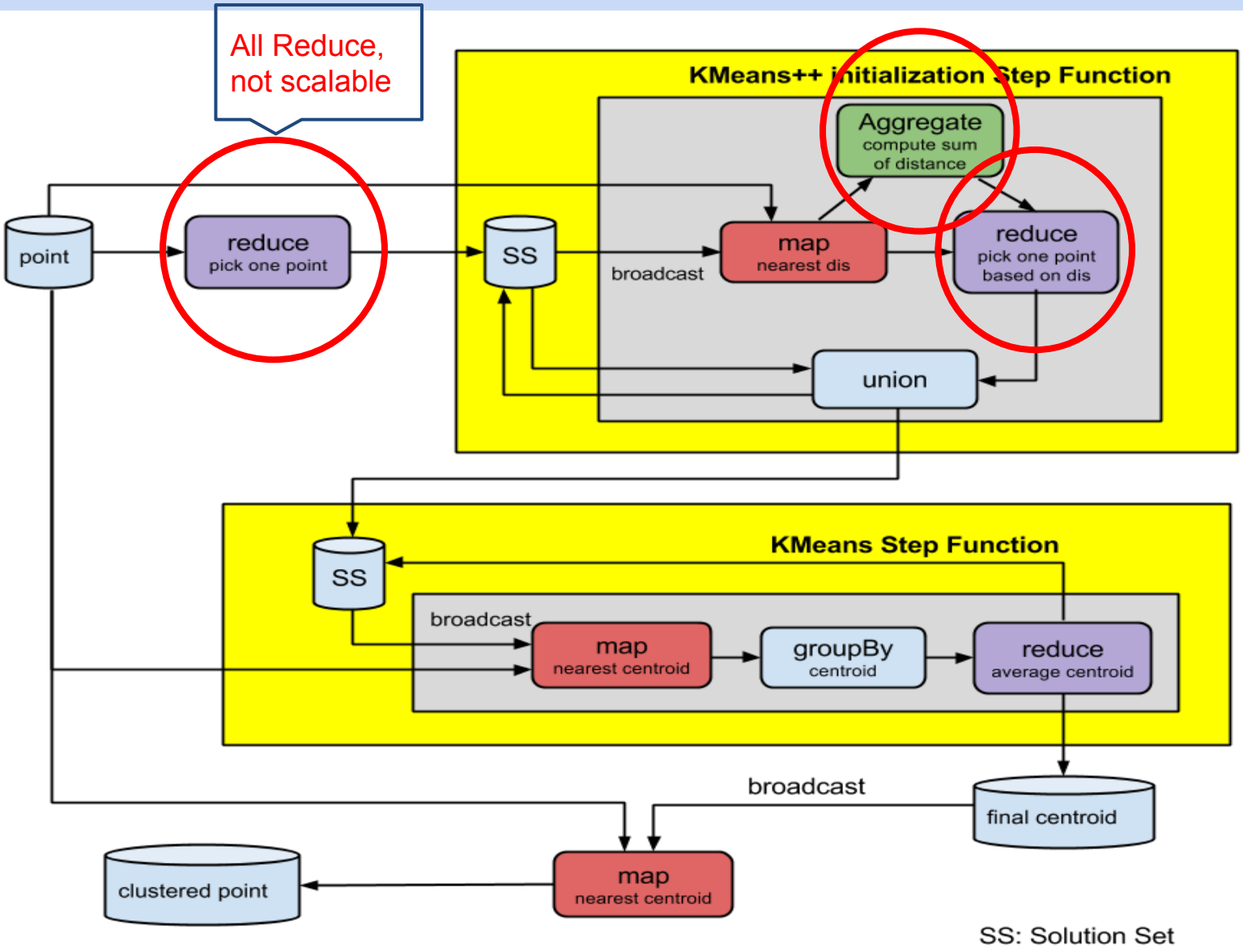
Experiments and Analysis



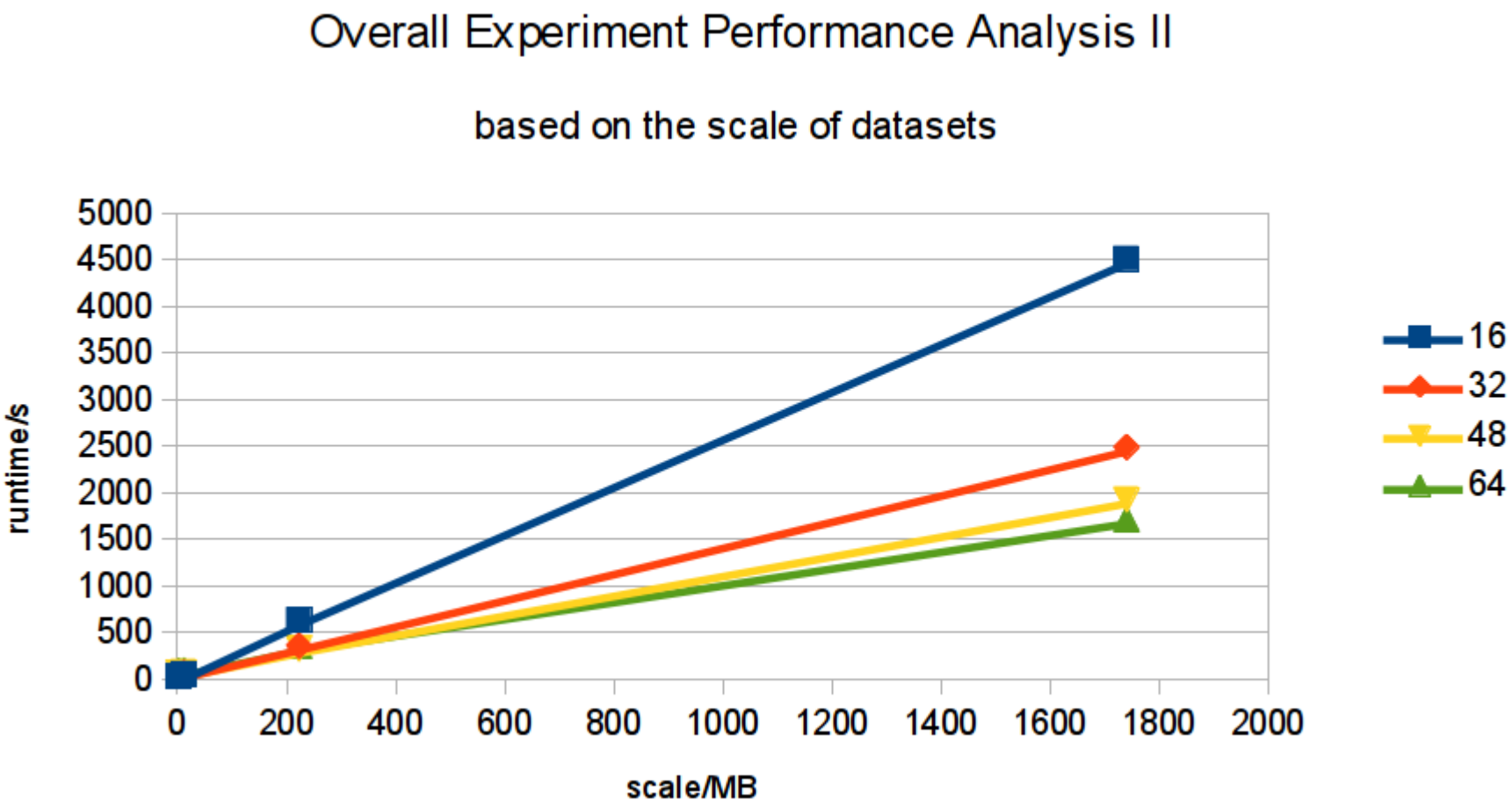
Experiments and Analysis

- Non-scalable Reduce Operations
- Network Transfer overhead
- Platform-dependent factors (internal implementation of task schedule etc.)

Experiments and Analysis



Experiments and Analysis



Thank you!