### INTRODUCTION

- **Motivation**
  - K-Medoids algorithm is one of the best-known clustering algorithm.
  - It is not widely used for big data analysis because of its high computational complexity.
  - Many studies have attempted to solve the problem.
  - However, all such studies have improved efficiency in the expense of accuracy.
  - We propose PAMAE, that achieves both high accuracy and high efficiency.

### OUR METHODOLOGY

- **Key Idea (PAMAE)**
  - Global search and entire data are the two main ingredients for high accuracy; but using them simultaneously harms efficiency.
  - To apply these two components individually though two phases:
    - **Phase I** is intended to quickly retrieve a high-quality set of seeds by sampling.
    - **Phase II** is intended to further reduce the possible errors induced by sampling.

### PROBLEMS OF EXISTING SOLUTIONS TO IMPROVE THE EFFICIENCY

- **Overall Procedure of PAMAE**
  - Phase I: Parallel Seeding
    - **Step (1):** Creates k random samples whose size is n.
    - **Step (2):** Runs a global search k-Medoids algorithm against each sample.
    - **Step (3):** Selects the best set of seeds among k sets of Medoids.
  - Phase II: Parallel Refinement
    - **Step (4):** Parition the entire data set just like the assignment step of the k-Medoids algorithm using the best seeds of Phase I.
    - **Step (5):** Update the Medoids of each cluster by choosing the most central object.
    - **Step (6):** Partitions the entire data set into clusters if needed.

### EVALUATION METHODOLOGY

- **Evaluation Results**
  - **Accuracy Result**
    - absolute error
  - **Relative Error Evaluation**
    - **Evaluation Results**

### EVALUATION RESULTS

- **Real-World Data Sets**
  - Data Set | # Object | # Time | Size | Type | Central
  - Cosmos2 | 2,494,293 | 36 | 7,998 | m | O
  - Covertype | 315,586,245 | 10 | 39,858 | True | O
  - TeraClickLog | 6,734,472,210 | 10 | 3,995,086 | True | O

- **Cornerstone of data is measured by Pearson's correlation coefficient of attributes.
  - Evaluation results of Cosmos2 are removed in the present due to lack of space.

- **Relative error evaluation**
  - **Efficiency Result**
    - **Scalability Test on Spark**
      - **Parallel Spark achieved near-linear scalability**
  - **Total elapsed time increased by 5.1 times when the data size increased from 30GB to 300GB by 10 times**

### CONCLUSION

- We proposed a novel parallel k-Medoids algorithm, which we call PAMAE.
- PAMAE consists of two phases: parallel seeding and parallel refinement.
- Our theoretical proof and experiment show that PAMAE achieved the excellent trade-off between accuracy and efficiency.
- We believe that our work has significantly raised the usability of the k-Medoids algorithm in the era of big data.

---

PAMAE: Parallel k-Medoids Clustering with High Accuracy and Efficiency
Hwanjun Song, Jae-Gil Lee* and Wook-Shin Han

Korea Advanced Institute of Science and Technology (KAIST) and Pohang University of Science and Technology (POSTECH)