



Dealing with Large Volumes of Data

Pertemuan 6

Contoh Implementasi:

- # Pseudocode MapReduce di Hadoop
- def map(key, value):
 - emit_intermediate(k1, v1)
- def reduce(key, values):
 - emit_final_result(k2, sum(values))

Parallel Multiple Classifier for Real-time Incremental Learning

Arsitektur Sistem:

- 1.Layer Input:** Data stream dari sumber heterogen
- 2.Distributed Processing Layer:**
 1. Pembagian tugas klasifikasi ke worker nodes
 2. Algoritma ensemble (Random Forest paralel)
- 3.Incremental Update:** Online learning dengan SGD

Keunggulan:

- Latensi rendah (94% faster vs. batch processing)
- Akurasi pertahankan 98.2% pada data drifting

Evaluasi Efektivitas PMCRI

Metrik	Formula	Target
Throughput	Data processed/second	$\geq 10\text{TB/jam}$
Scalability	$\text{Speedup} = T_1/T_n$ (n nodes)	Linear speedup
Accuracy	$(TP+TN)/(TP+TN+FP+FN)$	$>95\%$



Klasifikasi Inkremental

- **Algoritma Online Learning:**

1. **Stochastic Gradient Descent (SGD):**

- $w_{\{t+1\}} = w_t - \eta \nabla Q_i(w_t)$

1. Dimana η = learning rate, Q = loss function

2. **Naive Bayes Adaptif:**

1. Update prior probability secara dinamis

2. Cocok untuk data streaming (e.g., Twitter sentiment analysis)

- **Studi Kasus:**

- Sistem rekomendasi Netflix yang update model tiap 6 jam

Contoh Aplikasi

1. Deteksi Anomali Jaringan

- Pemrosesan 1M packets/detik menggunakan PMCRI
- Reduksi false positive 40%

2. Precision Agriculture

- Analisis real-time data drone + sensor IoT
- Prediksi hasil panen dengan akurasi 92%

Latihan

- **Kasus:**
Dataset e-commerce 50TB berisi:
- 1M transaksi/hari
- 200 fitur (user behavior, product stats)
- **Tugas:**
 1. Rancang arsitektur distribusi data (sharding strategy)
 2. Pilih algoritma klasifikasi inkremental
 3. Hitung throughput teoritis dengan 20 node (asumsi 1 node=5GB/detik)

Latihan

- **Referensi**

1. Dean, J. & Ghemawat, S. (2008). *MapReduce: Simplified Data Processing*. OSDI.
2. Bifet, A. (2010). *Adaptive Stream Mining*. Wiley.
3. UCI Repository (2023). Dataset untuk big data benchmarking.

- **Catatan:**

- Teknik paralelisasi dan inkremental learning penting untuk big data
- PMCRI terbukti efektif secara empiris

end

