 **DEPARTEMEN INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI DAN KOMUNIKASI**

**INSTITUT TEKNLOGI SEPULUH NOPEMBER**

**USULAN TUGAS AKHIR**

# IDENTITAS PENGUSUL

**NAMA : Setyassida Novian Putra Damara**

**NRP : 05111440000024**

**DOSEN WALI : Adhatus Solichah Ahmadiyah S.Kom, M.Sc**

**DOSEN PEMBIMBING : 1. Bagus Jati Santoso, S.Kom., Ph.D.**

**2. Royyana Muslim Ijtihadie S.Kom, M.Kom., Ph.D**

# JUDUL TUGAS AKHIR

“Desain dan Implementasi Aplikasi Pengolahan *Top-k Dominating Queries* Pada Data *Streaming* Terdistribusi”

# LATAR BELAKANG

Selama beberapa tahun terakhir ini, *preference queries*, kueri yang hanya mengambil *preferable* objek data dari *multidimensional* dataset mendapat perhatian lebih dari para peneliti yang mempunyai fokusan pada pengolahan basis data. Jenis kueri ini dapat memberikan berbagai macam hasil multi kriteria sehingga dapat dimanfaatkan dalam banyak aplikasi misalnya, *multimedia retrieval* [1]*, web search* [2], *analisis pasar* [3], dan e*-commerce* [3]. Dua metode yang paling banyak digunakan dalam *preference queries* adalah *top-k* dan *skyline queries*.

Misalnya, diberikan fungsi ranking (penilaian) : *d* → , di mana *d* adalah jumlah atribut (dimensi), kueri dengan metode *top-k* mengambil objek data *k* dengan nilai terbaik. Keuntungan dari metode *top-k queries* ini adalah pengguna dapat mengontrol jumlah objek data hasil dengan cara menetapkan parameter *k*. Kelemahan dari metode ini dirasa sulit bagi pengguna untuk menentukan peringkat yang sesuai fungsi.

Metode *skyline queries* mengatasi kelemahan pada metode *top-k queries* karena pada metode *skyline queries* tidak memerlukan fungsi peringkat apapun. Hasil kueri dari metode *skyline queries* terdiri dari objek data yang tidak didominasi oleh objek data lain dalam dataset tertentu. Hubungan dominasi didefinisikan sebagai berikut: misalnya diberi dua buah data *oi* dan *oj*, *oi* mendominasi *oj* jika *oi* tidak lebih buruk daripada semua atribut *oj* dan lebih baik dari pada *oj* setidaknya pada satu atribut. Dengan metode *skyline queries*, pengguna bisa mendapatkan objek data yang tidak lebih buruk dari data yang lain. Namun, pengguna tidak bisa mengendalikannya ukuran dari hasil, yang mungkin dapat mengembalikan hasil data yang sangat banyak.

Metode *top-k dominating queries* adalah metode kueri yang menggabungkan kelebihan dari metode *top-k* dan *skyline queries*. Lebih spesifik lagi, metode *top-k dominating queries* tidak memerlukan fungsi peringkat yang didefinisikan oleh pengguna dan dapat mengontrol ukuran hasil. Metode *top-k queries* yang mendominasi mengambil objek data *k* yang mendominasi jumlah tertinggi objek data dalam dataset tertentu. Artinya, nilai data objek *o* adalah jumlah objek data yang didominasi oleh *o*. Kueri teratas mendominasi identifikasi yang paling penting objek data secara intuitif. Hal ini dapat membantu banyak aplikasi seperti contoh di atas.

Ditambahin pengatntar tentan distributed environment

# RUMUSAN MASALAH

Rumusan masalah yang diangkat dalam tugas akhir ini dapat dipaparkan sebagai berikut:

1. Bagaimana cara mengolah *top-k dominating queries* pada data *streaming* terdistribusi?
2. Bagaimana merumuskan stuktur data yang terbaik untuk memecahkan permasalahan *top-k dominating queries* pada data *streaming* terdistribusi?
3. Bagaimana cara mengurangi biaya komputasi dan penyimpanan pengolahan t*op-k dominating queries* pada data *streaming* terdistribusi?

# BATASAN MASALAH

Permasalahan yang dibahas dalam tugas akhir ini memiliki batasan antara lain:

1. Nilai atribut yang akan diproses dalam algoritma ini bertipe numerik.
2. Tools yang digunakan dalam penelitian ini adalah Matlab.
3. *Dataset* yang digunakan berupa data *real-life* dan sintetis.
4. Konsep data *streaming* yang digunakan adalah *sliding windows* dengan tipe *count-based*.

# TUJUAN PEMBUATAN TUGAS AKHIR

Tujuan pembuatan tugas akhir ini antara lain:

1. Menemukan algoritma untuk mengolah *top-k dominating queries* pada data *streaming* terdistribusi.
2. Merumuskan struktur data yang terbaik untuk mendukung pemecahan permasalahan *top-k dominating queries* pada data *streaming* terdistribusi.
3. Merumuskan metode untuk mengurangi biaya komputasi dan penyimpanan *top-k dominating queries* pada data *streaming* terdistribusi.

# MANFAAT TUGAS AKHIR

Manfaat dari pembuatan tugas akhir ini adalah:

1. Meneliti permasalahan dalam pengolahan *top-k dominating queries* pada data *streaming* terdistribusi.
2. Mengusulkan dua algoritma/metode yang efisien untuk pengolahan *top-k dominating queries* pada data *streaming* terdistribusi, dengan harapan kedua algoritma ini mengurangi biaya komputasi secara signifikan.
3. Mengusulkan algoritma perkiraan dalam pengolahan *top-k dominating data computation*, dengan harapan dengan algoritma perkiraan ini dapat mengurangi biaya komputasi jika dibandingkan dengan *exact algorithm*.

# TINJAUAN PUSTAKA

## Top-K Queries

Untuk saat ini ada dua metode/teknik kueri berbasis referensi yang sangat sering digunakan, diantaranya adalah: (i) metode *top-k queries* dan (ii) metode *skyline queries*. Metode *top-k queries* menggunakan fungsi peringkat yang didefinisikan sebagai berikut , yang dimana akan memberikan nilai untuk setiap *tuple/points p* (*tuple* adalah sebuah objek yang merupakan kumpulan elemen hasil kueri). Hasil dari *top-k queries* terdiri dari *k* *points* yang memiliki nilai tertinggi sesuai dengan fungsi . Keunggulan dar metode/teknik ini adalah jumlah jawaban/keluaran yang dapat pengguna kontrol atau sesuaikan jumlahnya dengan meruabah parameter *k,* walaupun untuk beberapa kasus, kardinalitas dari jawaban/keluaran yang dihasilkan sama atau melebihi nilai parameter *k* (contohnya, terdapat duat atau lebih *points* yang memiliki nilai yang sama, dengan begitu keluaran yang dihasilkan dapat melebihi nilai *k* yang sudah ditetapkan.) Dalam beberapa kasus, semua nilai yang sama akan disertakan menjadi hasil atau depat juga dipangkas dengan kriteria tertentu sehingga jawaban/keluaran yang dihasilkan dapat sebanyak *k*. Salah satu keterbatasan metode ini adalah sangat bergantung pada fungsi peringkat, fungsi ini biasanya diatur oleh pengguna, yang dimana fungsi yang berbeda akan menghasikan jawaban/keluaran yang berbeda pula. Terlebih-lebih dalam beberapa kasus pemilihan fungsi yang sesuai tidak berdarkan intuitif. Contohnya dalam aplikasi *e-commerce*, tidak ada cara yang mudah dalam mengombinasikan atribut kecepatan *CPU* dan manajemen baterai untuk memilih laptop yang paling banyak diminati. [4]

## Skyline Queries

## Top-K Dominating Queries

## *Top-k dominating queries* merupakan gabungan antara dua *preference-based Query Processing* [2] yang paling banyak digunakan yaitu: 1) *Top-k query* dan 2) *skyline query*. *Top-k dominating queries* mengunakan fungsi ranking untuk merangking object (metode *Top-k queries*) dan mengunakan *dominating relationship* (metode *Skyline queries*). Tujuan dari metode *Top-k dominating queries* adalah mempertahankan keuntungan dan mengeliminasi keterbatasan dari metode *top-k queries* dan *skyline queries*. Maka dari itu *Top-k dominating queries* memiliki beberapa sifat khusus antara lain:

1. jumlah hasilnya terkontrol,

2. hasil tidak berubah,

3. tidak perlunya definisi penilaian dari user,

4. pada setiap objek akan memiki nilai yang akan menentukan rankingnya.

Top-K Dominating (TKD) query dapat didefinisikan sebagai sebuah query yang akan mengembalikan sebanyak K object yang memiliki score dominasi tertinggi pada suatu dataset yang diberikan. Penentuan score pada object dilakukan dengan cara menghitung jumlah object lain yang didominasi oleh object tersebut. Dalam hal ini yang dimaksut dengan dominasi adalah sebagai berikut. Diberikan dua buah object yaitu *o* dan *o’*. *o* dapat dikatan mendominasi *o’* yang disimbolkan dengan *o* ≺ *o'* jika dua kondisi ini terpenuhi. kondisi yang pertama pada setiap dimensi i pada object *o* tidak ada yang lebih besar dari dimensi *i* pada object *o’*. kondisi yang kedua terdapat minimal satu dimensi pada object *o* yang lebih kecil dari pada object *o’*.

## Sliding Windows

*Sliding Windows* adalah sebuah teknik dalam dunia pemrograman yang memodelkan data seperti sudut pandang kaca di dalam sebuah bis [3]. Terdapat *n* objek yang akan diamati dan *k* yang merupakan panjang sebuah kaca mengamatan atau dapat disebut juga panjang objek yang sedang diamat. Biasanya objek yang diamati disebut dengan objek aktif. Terdapat dua tipe *sliding window* yang pertama adalah *count based sliding window* dimana nilai *k* selalu konstan. Jika ada sebanyak *r* objek masuk menjadi objek aktif maka akan ada juga sebanyak *r* objek aktif yang akan kadaluarsa. Tipe yang kedua adalah *time-based sliding widow* dimana jumlah *k* tidak konstan. Pada tipe ini setiap objek memiliki waktu aktif masing-masing yang tidak saling terkaitan. Sehingga jumlah objek yang aktif bisa berbeda pada waktu yang berbeda.

## Grid Based Indexing

# RINGKASAN ISI TUGAS AKHIR

Tugas Akhir ini disusun untuk menangani masalah besarnya biaya komputasi dan biaya penyimpanan dalam pengolahan *top-k dominating queries* pada *data* *streaming* terdistribusi. Pada data ini saya penulis mengadopsi metode *count-based sliding window* [1], dimana dihasilkan objek data dalam *W* waktu dari waktu sekarang yang akan menjadi data pantauan.

Dalam proposal Tugas Akhir ini penulis mengusulkan dua metode pendekatan untuk *exact top-k dominating data monitoring*. Pertama adalah pendekatan berdasarkan penyaringan dimana dalam metode ini menggunakan data yang meiliki nilai dominasi tinggi sebagai penyaring untuk menhindari pengiriman data yang tidak perlu. Kedua dengan metode *cache-based* dengan harapan dapat mengurangi biaya komputasi secara signifikan. Dengan memanfaatkan kedua metode ini penulis mengusulkan sebuah algoritma untuk pengolahan *top-k dominating queries* pada data *streaming* terdistribusi. Untuk mengurangi biaya komunikasi dan komputasi, sebisa mungkin kita harus menghindari perhitungan yang berulang dari data *top-k dominating*.

Pada penelitian ini, penulis mengusulkan metode *lower*- and *upper-bounding* untuk memberikan nilai untuk tiap objek data yang ada, dengan menggunakian metode ini penulis akan secara cermat untuk memilih kandidat data yang diberikan dari total keseluruhan data. Penulis juga mengusulkan metode *sampling-based approximate* untuk perhitungan data *top-k dominating* dengan harapan dapat mengurangi biaya komunikasi dan komputasi secara signifikan dengan menjaga tingkat akurasi yang tinggi.

# METODOLOGI

## Penyusunan proposal tugas akhir

Proposal tugas akhir ini berisi tentang penjelasan mengenai pendahuluan dari tugas akhir yang akan dibuat. Pendahuluan ini terdiri dari hal yang melatarbelakangi tugas akhir, rumusan masalah yang diangkat, batasan masalah yang ada, tujuan dan manfaat dari tugas akhir ini. Selain itu, dijabarkan pula tinjauan pustaka yang digunakan sebagai referensi pendukung dalam pembuatan tugas akhir.

## Studi literatur

Pada studi literatur ini, akan dipelajari beberapa referensi yang akan diperlukan untuk membantu mendesain algoritma untuk mengolah *top-k dominating queries* pada data berbasis kelompok*.* Secara garis besar, ada tiga metode/teknik/algoritma yang akan menjadi pilar dalam tugas akhir kali ini. Yaitu, metode *top-k queries*, *skyline queries*, dan yang terakhir adalah gebungan dari kedua metode sebelumnnya dengan nama *Top-k Dominating Queries.*

## Analisis dan desain perangkat lunak

Pada tahap ini, penulis akan menganalisa masalah yang ada dalam pengolahan *top-k dominating query* pada data *streaming terdistribusi* dan mendesain algoritma yang dapat mengatasi permasalahan tersebut.

## Implementasi perangkat lunak

Perangkat keras yang digunakan adalah perangkat keras yang berbasis Windows demi kemudahan instalasi berbagai macam perangkat lunak yang dibutuhkan. Pada tugas akhir kali ini direncanakan penggunaan perangkat keras dan sistem operasi mempunyai spesifikasi sebagai berikut:

* + - Intel Core i3-2330M 2.2GHz
    - NVIDIA GeForce GT 920m 1GB
    - 4GB of RAM
    - Windows 10 64-bit

jika kemudian ternyata digunakan perangkat keras yang mempunyai spesifikasi yang berbeda, rincian spesifikasinya akan dilaporkan kemudian pada saat penyusunan laporan akhir pada tugas akhir kali ini.

Sedangkan untuk *environment* perangkat lunak, akan digunakan bahasa pemrograman Matlab dengan berbagai macam *library* pendukung.

## Pengujian dan evaluasi

Pengujian dilakukan untuk mengetahui apakah system yang dibangun dengan metode yang diusulkan telah bekerja dengan baik dan efisien atau belum. Pengujian dalam algoritma ini akan dilakukan dalam beberapa cara, antara lain:

1. Pengujian Akurasi
2. Pengujian Waktu Eksekusi
3. Pengujian Penggunaan Memori

## Penyusunan Buku Tugas Akhir

Pada tahap ini dilakukan penyusunan laporan yang menjelaskan dasar teori dan metode yang digunakan dalam tugas akhir ini serta hasil dari implementasi aplikasi perangkat lunak yang telah dibuat. Sistematika penulisan buku tugas akhir secara garis besar antara lain:

1. Pendahuluan
   1. Latar Belakang
   2. Rumusan Masalah
   3. Batasan Tugas Akhir
   4. Tujuan
   5. Metodologi
   6. Sistematika Penulisan
2. Tinjauan Pustaka
3. Desain dan Implementasi
4. Pengujian dan Evaluasi
5. Kesimpulan dan Saran
6. Daftar Pustaka

# JADWAL KEGIATAN

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tahapan** | **2017** | | | | **2018** | | | | | | | | | | | | | | | | | | | | |
| **Desember** | | | | **Januari** | | | | **Februari** | | | | **Maret** | | | | **April** | | | | **Mei** | | | | |
| Penyusunan Proposal |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |
| Studi Literatur |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |
| Perancangan Sistem |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |
| Implementasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |
| Pengujian dan Evaluasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |
| Penyusunan Buku |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | |  |  |  |  |

# DAFTAR PUSTAKA

|  |  |
| --- | --- |
| [1] | E. Tiakas, G. Valkans, A. N. Papadopoulos, Y. Manolopoulos and D. Gunopoulos, "Metric-based top-k dominating queries," *EDBT,* pp. 415-426, 2014. |
| [2] | D. Skoutas, D. Sacharidis, A. Simitsis, V. Kantere and T. Sellis, "Top-k dominant web services under multi-criteria matching," *EDBT,* pp. 898-909, 2009. |
| [3] | A. Yu, P. Agarwal and J. Yang, "Processing a large number of continuous preference top-k queries," *SIGMOD,* pp. 397-408, 2012. |
| [4] | M. Kontaki, A. N. Papadopoulos and Y. Manolopoulos, "Continuous Top-k Dominating Queries," *IEEE Transactions on Knowledge and Data Engineering,* vol. 24, no. 5, pp. 840 - 853, 2012. |