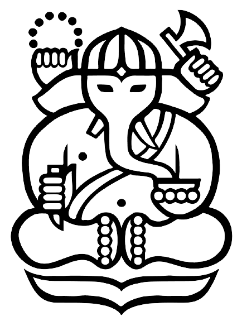
**PROPOSAL PROJECT**

**IF5270 SISTEM BIG DATA**

**Big Data Streaming menggunakan API Spotify untuk Rekomendasi Musik berdasarkan Model Clustering**

Disusun untuk memenuhi tugas mata kuliah Sistem Big Data



Disusun Oleh:

Iwan Kurniawan (23521058)

Hamzah Misbachul Adlan (23521070)

Brillian Fahmi Fauzi Perdana (23522024)

**INSTITUT TEKNOLOGI BANDUNG**

**SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA**

**PROGRAM STUDI INFORMATIKA**

**BANDUNG**

**2023**

**BAB I**

**PENDAHULUAN**

* 1. **Latar Belakang**

Dalam beberapa tahun terakhir, permintaan untuk rekomendasi musik yang personal telah semakin meningkat. Hal ini disebabkan oleh peningkatan jumlah pengguna layanan streaming musik seperti Spotify. Dalam hal ini, Big Data Streaming dapat dimanfaatkan untuk memberikan rekomendasi musik yang lebih akurat dan personal.

Dengan menggunakan data Spotify API yang tersedia, Big Data Streaming dapat menganalisis preferensi musik pengguna secara real-time dan memberikan rekomendasi musik yang relevan. Data yang diambil melalui Spotify API dapat dianalisis menggunakan berbagai teknik analisis data seperti machine learning, clustering, dan association rule mining.

Data Spotify API adalah data yang dapat diakses secara langsung dari Spotify, salah satu layanan streaming musik terbesar di dunia. Data Spotify API berisi informasi mengenai lagu, album, artis, playlist, dan sebagainya. Beberapa contoh data yang tersedia di Spotify API antara lain:

- Informasi mengenai lagu, seperti judul, artis, durasi, album, dan genre.

- Informasi mengenai artis, seperti nama, popularitas, dan genre.

- Informasi mengenai playlist, seperti judul, jumlah lagu, dan deskripsi.

* 1. **Persoalan**

Berdasarkan latar belakang yang telah disebutkan, waktu yang dihabiskan untuk listing data dari spotify membutuhkan waktu. Dalam tugas besar ini, teknologi big data seperti Hadoop, Spark, dan Apache Kafka dapat digunakan untuk memproses data Spotify API. Dengan menggunakan teknologi ini, data dari API Spotify dapat diolah dengan lebih efisien dan cepat.

Hadoop dapat digunakan untuk melakukan proses ETL (Extract, Transform, Load) pada data Spotify API, sehingga data tersebut dapat diolah lebih mudah dan cepat. Apache Spark dapat digunakan untuk melakukan proses analisis data dan machine learning dengan skala yang besar dan dalam waktu yang lebih cepat.

Selain itu, Apache Kafka dapat digunakan untuk mengambil data streaming secara real-time dari API Spotify. Dengan menggunakan Apache Kafka, data yang diambil dari API Spotify dapat diolah dan dianalisis secara real-time, sehingga dapat memberikan hasil yang lebih akurat dan tepat waktu.

* 1. **Kepentingan Persoalan**

Pentingnya persoalan ini diselesaikan adalah karena kebutuhan akan rekomendasi musik yang tepat sangat penting dalam dunia industri musik dan hiburan. Dalam beberapa industri, penggunaan teknologi big data dan machine learning dapat membantu meningkatkan pengalaman pengguna dengan memberikan rekomendasi musik yang sesuai dengan preferensi dan minat mereka.

Dengan memiliki sistem rekomendasi yang baik, platform musik dapat meningkatkan keterlibatan pengguna dan menarik lebih banyak pelanggan. Hal ini juga dapat meningkatkan loyalitas pelanggan dan memperluas basis pengguna dengan menawarkan rekomendasi musik yang lebih baik dan tepat sasaran.

* 1. **Tantangan**

Dalam Tugas Besar ini, terdapat beberapa tantangan sebagai berikut:

* Autentikasi: Untuk mengakses data Spotify, Kita harus memiliki akses token atau otentikasi. Token ini dapat berubah setiap saat, sehingga kita harus memperbarui token secara teratur agar tetap dapat mengakses data.
* Rate limiting: Spotify memiliki batasan tarif yang membatasi jumlah permintaan API yang dapat dilakukan dalam satu waktu. Jika kita melakukan permintaan terlalu sering atau terlalu banyak dalam waktu singkat, kita bisa dibatasi dan tidak dapat mengakses data selama beberapa waktu.
* Pengolahan Data yang Besar: Spotify memiliki miliaran lagu dan artis yang tersedia, dan memproses semua data tersebut dalam waktu yang singkat bisa sangat menantang. Oleh karena itu, memilih teknik pengolahan data yang tepat dan mengoptimalkan kinerja kode kita menjadi sangat penting.
* Perubahan dalam API: Spotify secara teratur memperbarui API mereka, sehingga kode yang telah ditulis sebelumnya mungkin tidak lagi berfungsi dengan benar. Kita harus memperbarui kode kita secara berkala untuk memastikan bahwa kita selalu dapat mengakses data terbaru.
* Penggunaan model Machile Learning : karena rekomendasi dibuat berdasarkan *clustering* diharapkan dapat dibuat model clustering yang cepat dan mampu mengelompokkan secara baik berdasarkan musik-musik yang telah didengarkan.
* Perlindungan Hak Cipta: Dalam mengambil data dari Spotify, kita perlu memastikan bahwa kita tidak melanggar hak cipta mereka. kita tidak dapat mengambil data tanpa izin atau memperbanyak data yang dilindungi hak cipta tanpa izin dari pemilik hak cipta.
  1. **Konsep Sistem**

Gambaran umum tentang konsep tool big data dengan data Spotify:

1. Pengambilan Data: Data Spotify dapat diambil dari API Spotify atau dengan cara web scraping menggunakan library seperti BeautifulSoup. Data dapat diambil dalam format JSON atau CSV.
2. Penyimpanan Data: Data Spotify kemudian disimpan dalam sistem penyimpanan big data, seperti Hadoop Distributed File System (HDFS), Apache Cassandra, MongoDB atau Amazon S3. Data yang disimpan dapat dianalisis secara cepat dan efisien menggunakan alat analisis data big data.
3. Proses Data: Data Spotify kemudian diproses menggunakan alat analisis big data, seperti Apache Spark atau Apache Flink. Alat-alat ini memungkinkan pengolahan data dalam skala besar dan real-time, yang memungkinkan pengolahan data Spotify secara efisien dan cepat.
4. Model Clustering: Model clustering dibuat untuk mengelompokkan lagu dan artis ke dalam kelompok-kelompok berdasarkan kesamaan karakteristik, seperti genre, tempo, dan lain-lain. Model clustering menggunakan algoritma machine learning, seperti K-Means atau Hierarchical Clustering, untuk menghasilkan kelompok-kelompok yang optimal.
5. Visualisasi Data: Setelah model clustering terbentuk, sistem akan memberikan rekomendasi musik berdasarkan kelompok-kelompok yang terbentuk. Rekomendasi musik ini dapat diberikan kepada pengguna melalui antarmuka aplikasi web atau aplikasi mobile yang terhubung dengan sistem.
   1. **Infrastruktur**

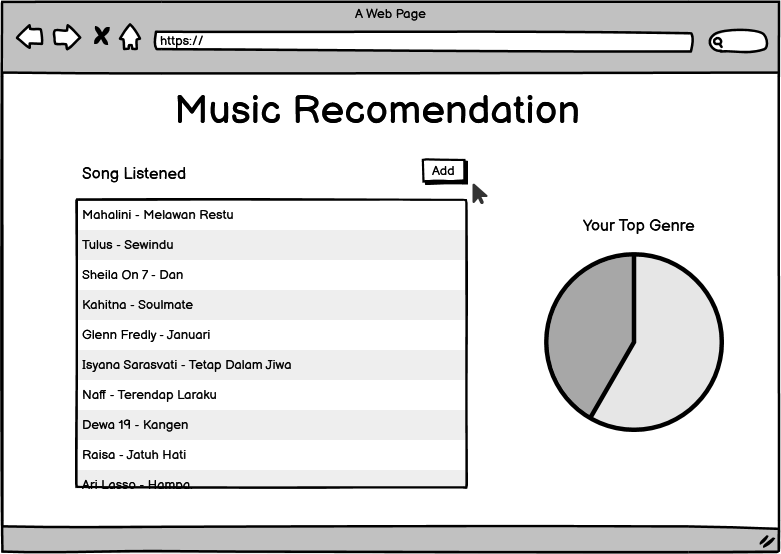
****

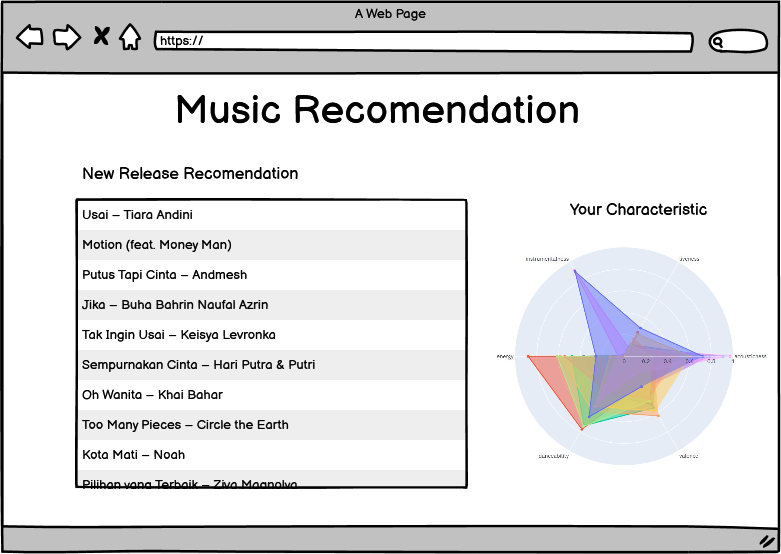
Dalam konteks Big Data, ada beberapa alat atau tool yang dapat digunakan untuk memproses data Spotify. Beberapa tool big data yang rencananya akan digunakan dalam pengolahan data Spotify adalah sebagai berikut:

1. Apache Hadoop: Hadoop adalah *framework open source* yang digunakan untuk memproses data besar. Hadoop terdiri dari beberapa komponen, seperti HDFS (Hadoop Distributed File System) dan MapReduce, yang memungkinkan pengolahan data dalam skala besar secara terdistribusi. Hadoop dapat digunakan untuk memproses data Spotify dan melakukan analisis data seperti pengolahan log dan profil pengguna.
2. Apache Kafka: Seperti yang telah disebutkan sebelumnya, Kafka adalah platform pengolahan streaming data yang handal dan realtime. Kafka dapat digunakan untuk memproses data Spotify dalam waktu nyata, sehingga data dapat diakses dan digunakan dengan cepat.
3. Apache Spark: Spark adalah kerangka kerja pengolahan data terdistribusi yang digunakan untuk memproses data besar dengan cepat. Spark mendukung pengolahan data dalam skala besar dan real-time, yang memungkinkan untuk memproses data Spotify secara cepat dan efisien.
4. MongoDB: Database NoSQL open-source yang digunakan untuk menyimpan data dengan skema fleksibel dan skalabilitas yang tinggi.
5. Apache Hive: Hive adalah sebuah data warehouse yang digunakan untuk query dan analisis data yang besar dan kompleks dalam ekosistem Hadoop. Hive menggunakan bahasa SQL untuk melakukan query data, sehingga memudahkan pengguna yang sudah terbiasa dengan SQL.
6. Spark ML: Library untuk machine learning pada Apache Spark yang digunakan untuk membuat model machine learning seperti clustering, klasifikasi, dan regression.
7. Laravel: Framework PHP open-source yang digunakan untuk membangun aplikasi web dan API. Laravel menyediakan fitur-fitur yang lengkap seperti routing, template engine, dan ORM (Object-Relational Mapping) untuk mempermudah pembuatan aplikasi web.

Kombinasi dari tool dan teknologi tersebut dapat digunakan untuk membangun infrastruktur yang handal dan efisien dalam memproses data dan menghasilkan rekomendasi musik berdasarkan data dari Spotify API.

* 1. **Gambaran Ouput**

****

****

* 1. **Peran**

Berikut adalah pembagian tugas yang akan dilakukan dalam membangun sistem rekomendasi musik berdasarkan data Spotify dengan menggunakan teknologi big data dan machine learning:

1. Analisis kebutuhan dan desain sistem:

* Menganalisis kebutuhan dan spesifikasi sistem.
* Mendesain arsitektur sistem dan skema database.

1. Pengambilan data dari API Spotify:

* Menerapkan autentikasi dengan menggunakan OAuth2.
* Mengambil data dari API Spotify dengan menggunakan Python dan Kafka.

1. Data preprocessing dan pipeline:

* Membersihkan dan mengubah format data Spotify menjadi format yang sesuai untuk pemrosesan di Spark.
* Menyimpan data hasil preprocessing di MongoDB.
* Menyediakan pipeline data

1. Pemodelan dan clustering:

* Membuat model machine learning untuk clustering menggunakan Spark ML.
* Menentukan jumlah cluster yang optimal.
* Mengelompokkan data berdasarkan cluster yang telah ditentukan.
* Pengembangan aplikasi web:

1. Membangun aplikasi web dengan framework Laravel.

* Menghubungkan aplikasi web dengan MongoDB untuk menampilkan hasil clustering.
* Menyediakan fitur untuk merekomendasikan lagu berdasarkan lagu lagu yang dipilih.

1. Uji coba dan evaluasi:

* Melakukan uji coba sistem dan mengevaluasi performa serta akurasi sistem.
* Mengoptimalkan performa sistem dengan mempertimbangkan faktor-faktor seperti latency dan throughput.

1. Dokumentasi:

* Mendokumentasikan seluruh proses pengembangan sistem, mulai dari tahap analisis kebutuhan hingga tahap evaluasi dan optimasi.
* Menyediakan dokumentasi penggunaan sistem untuk memudahkan pengguna dalam menggunakan sistem.

Berikut peran dari masing masing anggota tim :

| No | Nama | Peran |
| --- | --- | --- |
| 1 | Iwan Kurniawan | Team Leader, Documenter |
| 2 | Hamzah Misbachul Adlan | ML Engineer |
| 3 | Brillian Fahmi Fauzi Perdana | Data Engineer, Web Developer |

Selain pembagian peran, juga dibutuhkan penunjukan Person In Charge (PIC) untuk setiap tugas yang telah dibagi. Berikut adalah pembagian PIC untuk masing-masing tugas:

| Nomor | Tugas | PIC |
| --- | --- | --- |
| 1 | Analisis kebutuhan dan desain sistem | Iwan Kurniawan |
| 2 | Pengambilan data dari API Spotify | Brillian Fahmi Fauzi Perdana |
| 3 | Data preprocessing dan pipeline | Hamzah Misbachul Adlan |
| 4 | Pemodelan dan clustering | Hamzah Misbachul Adlan |
| 5 | Membangun aplikasi web dengan framework Laravel | Brillian Fahmi Fauzi Perdana |
| 6 | Uji coba dan evaluasi | Iwan Kurniawan |
| 7 | Dokumentasi | Iwan Kurniawan |

* 1. **Metric Evaluasi**

Sebagai sistem Big Data untuk rekomendasi musik, terdapat beberapa metrik evaluasi yang akan digunakan, antara lain:

Kecepatan: Metrik ini mengukur seberapa cepat sistem dapat memberikan rekomendasi kepada pengguna. Semakin cepat sistem memberikan rekomendasi, semakin baik karena pengguna tidak perlu menunggu terlalu lama.

Presisi: Metrik ini mengukur seberapa akurat rekomendasi yang diberikan oleh sistem. Semakin akurat sistem memberikan rekomendasi, semakin baik karena pengguna akan lebih puas dengan rekomendasi yang diberikan.

Untuk mengukur kecepatan, dapat digunakan metrik seperti waktu respon atau throughput. Sedangkan untuk mengukur presisi, dapat digunakan metrik seperti recall atau precision.