SEPAKAT - Modul Integrasi: Integrasi Data Regsosek, sebagai Informasi Dasar Individu, dengan Data Terkait menggunakan Pendekatan Resolusi Entitas

Pendahuluan

Latar Belakang

Penanggulangan kemiskinan merupakan isu prioritas dalam beberapa tahun terakhir dan telah disahkan dalam dokumen Rencana Pembangunan Jangka Menengah Nasional (RPJMN) dan Daerah (RPJMD) periode terkini. Berbagai upaya terus dilakukan untuk menjawab tantangan tersebut, termasuk diantaranya dengan mewujudkan reformasi sistem perlindungan sosial melalui penyediaan data sosial ekonomi yang terintegrasi dan akurat, serta mencakup setiap individu di seluruh wilayah Indonesia. Pengumpulan data Registrasi Sosial Ekonomi (Regsosek) merupakan salah satu upaya mencapai perwujudan reformasi tersebut. Regsosek diharapkan dapat menghasilkan data terpadu yang tak hanya dimanfaatkan untuk program perlindungan sosial, melainkan turut mampu mencakup seluruh program yang dibutuhkan masyarakat kedepannya.

Regsosek diproyeksikan turut dapat digunakan untuk kepentingan perencanaan dan evaluasi pembangunan, serta memastikan peruntukan kebijakan pemerintah yang lebih terarah. Regsosek akan hadir sebagai penyedia informasi dasar tentang kondisi sosial ekonomi di level individu. Data tersebut kemudian akan digunakan sebagai acuan target penerima program dan bantuan pemerintah sebagai upaya penanggulangan kemiskinan. Selain, tidak menutup kemungkinan untuk diintegrasikan dengan data sektoral yang bersesuaian demi tercapainya tujuan program dan bantuan tersebut. Namun sejauh ini beberapa program dan bantuan tersebut masih dikelola secara individu oleh masing-masing kementerian dan atau lembaga. Hal ini menjadi tantangan besar bagi Regsosek untuk tetap berdiri sebagai acuan data tunggal, meskipun diintegrasikan ataupun digunakan berbagi-pakai antar kementerian dan lembaga. Regsosek turut ditantang untuk mampu menjembatani koordinasi berbagi-pakai data hingga lintas daerah yang tetap memastikan pemakaian data individu yang konsisten. Proses tersebut menjadi bagian penting dalam perjalanan memaksimalkan potensi Regsosek mendatang, serta perlu direncanakan sejak dini dan matang.

Dalam rancangan jangka pendek - menengah, Data Regsosek akan dimanfaatkan untuk meningkatkan kualitas berbagai layanan pemerintah, seperti pendidikan, bantuan sosial, kesehatan, dan administrasi kependudukan. Data Regsosek juga akan dapat digunakan untuk mengidentifikasi kebutuhan kelompok masyarakat yang paling rentan ketika menghadapi bencana, wabah, atau kondisi sosial ekonomi lainnya. Misalnya kelompok perempuan yang berperan sebagai KRT (kepala rumah tangga), disabilitas, kelompok penduduk berpenyakit kronis, kaum lansia, dan anak-anak. Sedangkan untuk jangka panjang, Regsosek bertujuan untuk membangun data kependudukan tunggal yang berkualitas, mutakhir, dan akurat, sehingga dapat dimanfaatkan oleh pemerintah untuk menyusun dan menjalankan kebijakan secara tepat sasaran.

Secara teknis, integrasi data dapat didefinisikan sebagai proses menggabungkan data dari berbagai sumber menjadi satu kesatuan yang koheren. Sebagai contoh konkrit kedepan, seperti tertuang dalam *roadmap* pelaksanaan Regsosek di tahun 2023, salah satu tantangan terbesar yang akan dihadapi untuk memaksimalkan potensi Regsosek adalah mekanisme integrasi dan sinkronisasi dengan sumber data seperti Percepatan Penghapusan Kemiskinan Ekstrem (P3KE) dan Data Terpadu Kesejahteraan Sosial (DTKS). Meski telah direncanakan akan diintegrasikan dengan Nomor Induk Kependudukan (NIK), namun banyak kemungkinan dapat terjadi dalam proses integrasi ini yang membuka peluang terjadinya kekeliruan hasil sinkronisasi. Seperti diketahui, dari hasil penelusuran pada sampel data awal, masih ditemui kegagalan proses sinkronisasi pengidentifikasian kecocokan individu antar data yang hanya mendasarkan pada data NIK saja. Sedangkan, kesatuan yang koheren adalah tujuan utama pengumpulan data Regsosek, yang tentunya dapat menunjang tercapainya pemahaman menyeluruh tentang kondisi masyarakat dari berbagai aspek dan sudut pandang yang dihasilkan dari proses integrasi tersebut. Hal ini sejalan dengan tujuan lain Regsosek yaitu mampu menjadi landasan dalam penyusunan kebijakan yang komprehensif berbasis bukti (*evidence-based policy*).

Sejalan dengan itu, sejatinya sejak 2016 telah dikembangkan sistem yang mendukung pelaksanaan Perencanaan, Penganggaran, Pemantauan, Evaluasi dan Analisis Kemiskinan Terpadu yang bernama SEPAKAT. Sistem aplikasi ini dirancang dalam rangka pelaksanaan strategi program secara holistik, tematik, integratif, dan spasial. SEPAKAT telah berhasil berkontribusi dalam proses perencanaan dan penganggaran daerah berbasis bukti dan data sehingga dapat secara efektif merumuskan program kemiskinan secara tepat sasaran. Sebagai basis perencanaan, SEPAKAT tersedia sebagai perangkat analisis yang turut memuat berbagai data yang utamanya berkaitan dengan kemiskinan, termasuk diantaranya data olahan setiap sektor untuk masing-masing daerah.

Kedepannya, Regsosek akan hadir sebagai informasi dasar kondisi sosial ekonomi dan akan memegang peranan penting sebagai basis perencanaan terkini percepatan penanggulangan kemiskinan. Tak dapat dipungkiri bahwa integrasi dengan SEPAKAT adalah sebuah keniscayaan, dan isu identitas akan bertindak sebagai tantangan utama untuk mencapai kesempurnaan integrasi ini. Seperti diketahui, SEPAKAT sudah berhasil mengintegrasikan berbagai sumber data sektoral, namun apabila berkaitan dengan proses pencocokan data hingga level individu, SEPAKAT hanya mendasari integrasi menggunakan kolom NIK, dan terbukti (pada data sampel) masih banyak menghasilkan ketidakcocokan hasil sinkronisasi.

Tantangan Integrasi Data Regsosek

Resolusi entitas (*entity resolution*), juga dikenal sebagai pencocokan entitas atau *record linkage*, adalah proses mengidentifikasi dan menautkan entitas yang sesuai dengan antara satu data dengan data lainnya yang merujuk ke satu entitas yang sama. Dalam konteks Regsosek, resolusi entitas digunakan untuk mengidentifikasi dan menautkan data yang merepresentasikan individu penduduk ataupun rumah tangga dan keluarga yang sama dari berbagai kumpulan data yang berbeda. Hal ini dapat dicapai dengan menggunakan kombinasi teknik seperti pencocokan data (*data matching)*, deduplikasi data (*data deduplication)*, dan pembersihan data (*data cleaning)*.

Data Regsosek dijadwalkan akan mulai tersedia pada pertengahan tahun 2023 dan diproyeksikan akan digunakan sebagai rujukan pelaksanaan berbagai program pemerintah, termasuk diantaranya adalah penghapusan kemiskinan ekstrem. Data Regsosek yang dikumpulkan secara sensus pastinya akan berukuran sangat besar dan berisi ratusan juta baris data, sehingga integrasi dan sinkronisasi data secara manual akan menjadi tidak praktis dan menghabiskan waktu. Oleh karena itu, metode resolusi entitas yang terautomasi dianggap perlu untuk dipertimbangkan sebagai solusi dalam mengidentifikasi dan menautkan individu antar data secara lebih efisien, serta lebih mudah *(scalable)* untuk diterapkan bahkan ke skala nasional.

Salah satu metode yang dapat digunakan untuk resolusi entitas pada data Regsosek adalah dengan menggunakan metode deterministik (*deterministic linkage*), yang melibatkan penerapan seperangkat aturan untuk menentukan apakah dua atau lebih entitas merujuk kepada entitas yang sama. Contohnya adalah dengan melakukan pencocokan data berdasarkan NIK sebagai identitas unik dari setiap penduduk.

Keunggulan dari metode deterministik adalah dapat dilakukan dengan mudah dan cepat serta memiliki tingkat keakuratan yang tinggi karena menggunakan informasi pengenal pribadi, sehingga kecil kemungkinan untuk menghubungkan data dari individu yang berbeda. Namun demikian, metode ini menjadi kurang fleksibel karena pada sejumlah kasus data identitas unik seperti NIK bisa saja memiliki tingkat kesalahan pencatatan yang tinggi, terlebih data Regsosek dikumpulkan secara masif melalui mekanisme pengumpulan data langsung dari lapangan oleh petugas dengan tingkat kualitas pengumpulan data yang sangat beragam, sehingga rentan akan terjadinya data hilang (*missing value*), data inkonsisten dan salah eja (*typo*), serta berbagai masalah lainnya yang menyebabkan metode deterministik tidak dapat digunakan.

Probabilistic Record Linkage **sebagai Solusi Integrasi Data**

Metode lainnya yang dapat menjadi alternatif untuk resolusi entitas pada data Regsosek adalah menggunakan metode probabilistik (*probabilistic linkage*), yaitu penggunaan ukuran kemiripan nilai atribut/karakteristik (*attribute similarity*) yang tersedia pada data. Sebagai contoh, nilai atribut yang dimaksud dapat berupa nama lengkap individu, nama ibu kandung, tanggal lahir, jenis kelamin, alamat, dan sebagainya, yang dapat dibandingkan dan digunakan untuk menghitung probabilitas bahwa dua entitas merujuk pada individu yang sama. Metode probabilistik dapat digunakan menjadi alternatif solusi untuk integrasi dan sinkronisasi data Regsosek yang menawarkan sejumlah keunggulan sebagai berikut:

- Dapat melakukan pentautan (*linkage*) data dengan informasi parsial atau tidak utuh, sehingga dapat meningkatkan cakupan integrasi data.
- Dapat menangani kesalahan dan ketidakkonsistenan dalam data, seperti salah eja atau variasi ejaan.
- Memungkinkan penggabungan informasi dan sumber data tambahan, seperti data demografis ataupun data lokasi (*geocoding*), untuk meningkatkan akurasi pencocokan data.

Meskipun metode probabilistik tidak menjamin akurasi pentautan data secara sempurna, namun kajian ilmiah dari Zhu, dkk. (2015)[1] menyimpulkan bahwa metode probabilistik secara umum mengungguli metode deterministik karena kemampuannya untuk menyeimbangkan *trade-off* antara sensitivitas dan presisi penentuan *matching* terlepas dari beragamnya kualitas data yang perlu diintegrasikan.

Berdasarkan uraian permasalahan dan kajian studi diatas, dapat disimpulkan bahwa perlu adanya sebuah modul khusus yang mampu mempersiapkan proses integrasi Regsosek, sebagai informasi dasar, dengan sumber data lain terkait, yaitu modul Integrasi. Kami meyakini bahwa integrasi dan sinkronisasi data antar lembaga pemerintah dan berbagai sektor dengan memanfaatkan Regsosek sebagai data dasar merupakan kunci utama untuk dapat mewujudkan visi **membangun Indonesia dengan data**. Berkenaan dengan itu, proposal ini akan membahas usulan modul Integrasi untuk *platform* SEPAKAT, yang akan memperkaya modul yang telah dibangun sebelumnya. Adapun, proses pembangunan modul Integrasi ini diperlukan agar direncanakan dengan matang dan diteliti sejak dini, mengingat berbagai potensi permasalahan sebagai berikut:

- Sebelumnya belum dipersiapkan rencana integrasi sinkronisasi antar data selain menggunakan NIK;
- Pemrosesan integrasi dan sinkronisasi data di level individu yang notabene berskala besar baik dalam ukuran, muatan, ataupun jumlah baris data;
- Pemrosesan integrasi dengan data lama yang belum dapat dipastikan keutuhannya;
- Ragam *non-sampling error* penyebab terhambatnya proses pentatutan data individu dalam proses integrasi;
- Duplikasi data individu pada Regsosek.

Tujuan

Tujuan utama dari proposal ini adalah untuk menyediakan solusi modul pemrosesan integrasi data Regsosek dengan berbagai sumber data terkait, terutama pada level individu, baik terhadap data terdahulu maupun sebagai persiapan integrasi dengan pendataan mendatang. Secara lebih rinci tujuan lainnya adalah sebagai berikut:

- Dengan metode pendataan menyeluruh (sensus), solusi modul pemrosesan integrasi ini bertujuan untuk mempertahankan Regsosek sebagai informasi dasar acuan untuk data lainnya, yang menghasilkan data tunggal individu yang konsisten.
- Mempersiapkan wadah perluasan data dan informasi Regsosek, sebagai informasi dasar acuan, hasil integrasi dan sinkronisasi dengan sumber data lain, dengan turut mempertahankan keutuhan koheren data tunggal individu yang konsisten.

[1] Zhu, Ying, et al. "When to conduct probabilistic linkage vs. deterministic linkage? A simulation study." *Journal of biomedical informatics* 56 (2015): 80-86. https://doi.org/10.1016/j.jbi.2015.05.012

- Menyediakan modul Integrasi bagi SEPAKAT yang dapat memperkaya modul yang telah ada saat ini yaitu modul Analisis, Perencanaan, Penganggaran, Monitor, dan Evaluasi.
- Menyediakan modul Integrasi bagi SEPAKAT yang mampu beradaptasi dengan pemrosesan bersama data besar berskala nasional.
- Memastikan *one single ID* untuk setiap individu hasil pemrosesan integrasi dengan Regsosek sebagai informasi dasar acuan.
- Merumuskan formulasi kombinasi atribut penentu tautan individu antar data.

Manfaat

Secara garis besar, manfaat dari solusi yang diusulkan adalah untuk menjadi *support system* yang terhubung dengan *platform* SEPAKAT, dengan fungsi utama sebagai *high scalable integrator system* berbasis *vector database* yang memudahkan proses integrasi data Regsosek dengan berbagai sumber data lainnya seperti DTKS dan P3KE menggunakan pendekatan *probabilistic record linkage*.

Dengan demikian, modul ini dapat bermanfaat untuk membantu pemrosesan integrasi data Regsosek secara modern dan robust untuk selanjutnya dapat digunakan oleh *platform* SEPAKAT sebagai dasar pembuatan analisis dan perencanaan anggaran, memungkinkan analis dan pemangku kepentingan untuk melihat data dari berbagai sudut pandang dan menemukan pola atau hubungan antar data yang mungkin tidak terlihat jika data tersebut tetap terpisah. Hal ini dapat membantu dalam merumuskan keputusan yang lebih tepat dan bermanfaat. Selain itu, integrasi data juga dapat membantu dalam mengelola data secara lebih efisien karena data yang terintegrasi dapat diakses dengan mudah dan cepat.

Ide Gagasan

Fitur Utama

Aplikasi yang diusulkan merupakan sebuah aplikasi pemrosesan data yang dapat digunakan untuk mengintegrasikan data dari berbagai sumber. Aplikasi ini dapat menampilkan data-data yang saling terkait dari berbagai sumber data lintas instansi, kementerian/lembaga secara akurat dan komprehensif. Aplikasi ini sangat bermanfaat untuk kepentingan perencanaan program nasional, yang memiliki unit sasaran pelaksanaan pada tingkat kabupaten, desa, keluarga, hingga individu.

Data yang digunakan dalam perencanaan suatu kebijakan atau penyusunan program, sedianya dikumpulkan dari berbagai macam sumber, dan memiliki struktur serta konsep dan definisi yang berbeda. Di sisi lain, Untuk menetapkan suatu kebijakan atau melaksanakan suatu program, data yang akurat dan komprehensif diperlukan sebagai dasar pertimbangan. Oleh karena itu kami mengusulkan aplikasi untuk pengintegrasian data yang memiliki fitur berupa :

- **Pemrosesan pembersihan data dari duplikasi (Deduplikasi)**
  Fitur ini berguna untuk mengetahui potensi duplikasi *record* dari data-data yang tersedia. Pengguna dapat menginputkan wilayah yang akan diteliti, kemudian aplikasi akan menampilkan daftar *record* yang berpotensi menjadi duplikat dari *record-record* data pada wilayah tersebut. Selanjutnya, berdasarkan data tersebut, pengguna dapat melakukan konfirmasi atau rekonsiliasi agar duplikasi data dapat dihindari.

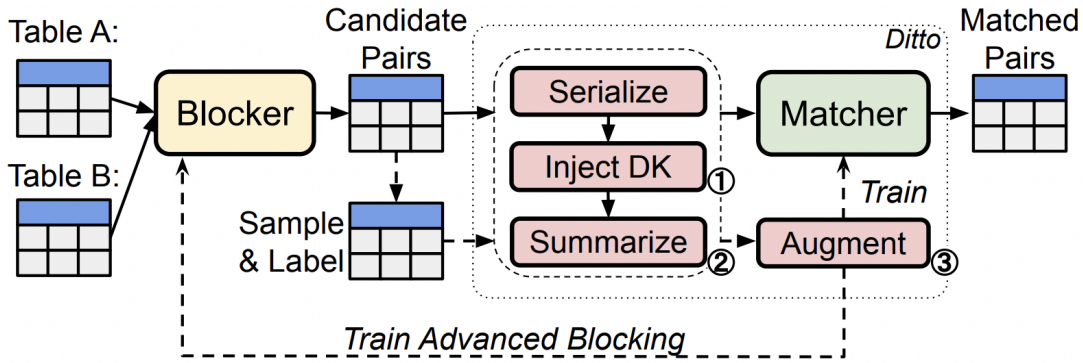
- **Pentautan (*linkage*) dari data satu ke data-data lainnya.**
  Fitur ini berfungsi untuk menampilkan *record* data menurut wilayah, dan tautannya (*linkage*) dengan data-data lain. Hal ini sangat bermanfaat untuk melihat data secara komprehensif, bukan dari satu data saja namun variabel-variabel lain pada data-data yang tersedia. Pengguna dapat memilih data dasar sebagai basis pencarian, kemudian aplikasi akan menyajikan tautan (*linkage*) dari semua data yang tersedia berdasarkan data dasar yang dipilih tersebut.

- **Pemeriksaan dan kontrol kualitas terhadap single ID yang digunakan saat ini (NIK).**
  Pengguna dapat melakukan pencarian menggunakan NIK pada semua data yang tersedia, kemudian aplikasi akan menampilkan record-record yang memiliki NIK tersebut. Setelah itu, akan ditampilkan pula skor *linkage* record-record tersebut. Apabila score tersebut tinggi maka dapat disimpulkan pencarian dengan NIK sudah konsisten diantara data-data yang tersedia, sebaliknya, jika rendah maka ada potensi kesalahan konten pada record data tersebut. Jika skor *linkage* rendah, aplikasi akan menampilkan data dengan skor *linkage* yang paling baik dengan data tersebut.

- **Pencarian secara realtime data komprehensif dari individu.**
  Pada fitur ini pengguna dapat melakukan pencarian secara realtime data individu dengan menginputkan nama, umur, jenis kelamin, dan nama kepala keluarga nya. Kemudian aplikasi akan melakukan pencarian dan menampilkan hasilnya kepada pengguna. Data yang ditampilkan bersumber dari semua data yang telah terindex di dalam sistem.

Proses Bisnis

Alur kerja utama dari sistem pentautan pada aplikasi ini dapat diuraikan melalui Gambar 1 di bawah:



Gambar 1. Alur kerja utama proses pentautan individu antar sumber data

Pendekatan alur kerja tersebut disadur dari penelitian oleh Li, dkk. (2020)[2] yang telah melakukan penelitian serupa. Berikut penjelasan rinci alur kerja utama proses pentautan individu antar sumber data pada Gambar 1.

1. Sumber data menyediakan input data untuk penyelesaian entitas, yang dapat berupa representasi kalimat entitas, seperti orang atau rumah tangga. Misalnya dalam penggunaan *linkage*, data akan dikumpulkan baik dari Regsosek maupun DTKS.

2. Kalimat tersebut akan digunakan untuk menyempurnakan model SentenceBERT, sehingga kami dapat menghasilkan representasi berkualitas tinggi dari setiap entitas ke dalam format embedding. Hal ini dapat dilakukan dengan memilih sampel secara acak kemudian secara manual membuat beberapa pasangan yang cocok/*match* dari Regsosek dan DTKS sebagai *labelled train data*.

3. Model SentenceBERT yang telah di tunning dengan baik akan digunakan untuk menghasilkan embedding dari setiap entitas yang tersedia di dalam data. Setelah itu, hasil embedding akan dimasukkan ke dalam database Milvus.

4. Langkah berikutnya adalah *Blocking*. Untuk setiap entitas, sistem akan melakukan pencarian vektor di Milvus berdasarkan *cosine similarity* untuk menemukan kandidat entitas top-K dari database dengan skor kesamaan tertinggi. Dari proses ini kita akan mendapatkan semua kandidat pasangan yang mungkin cocok/match.

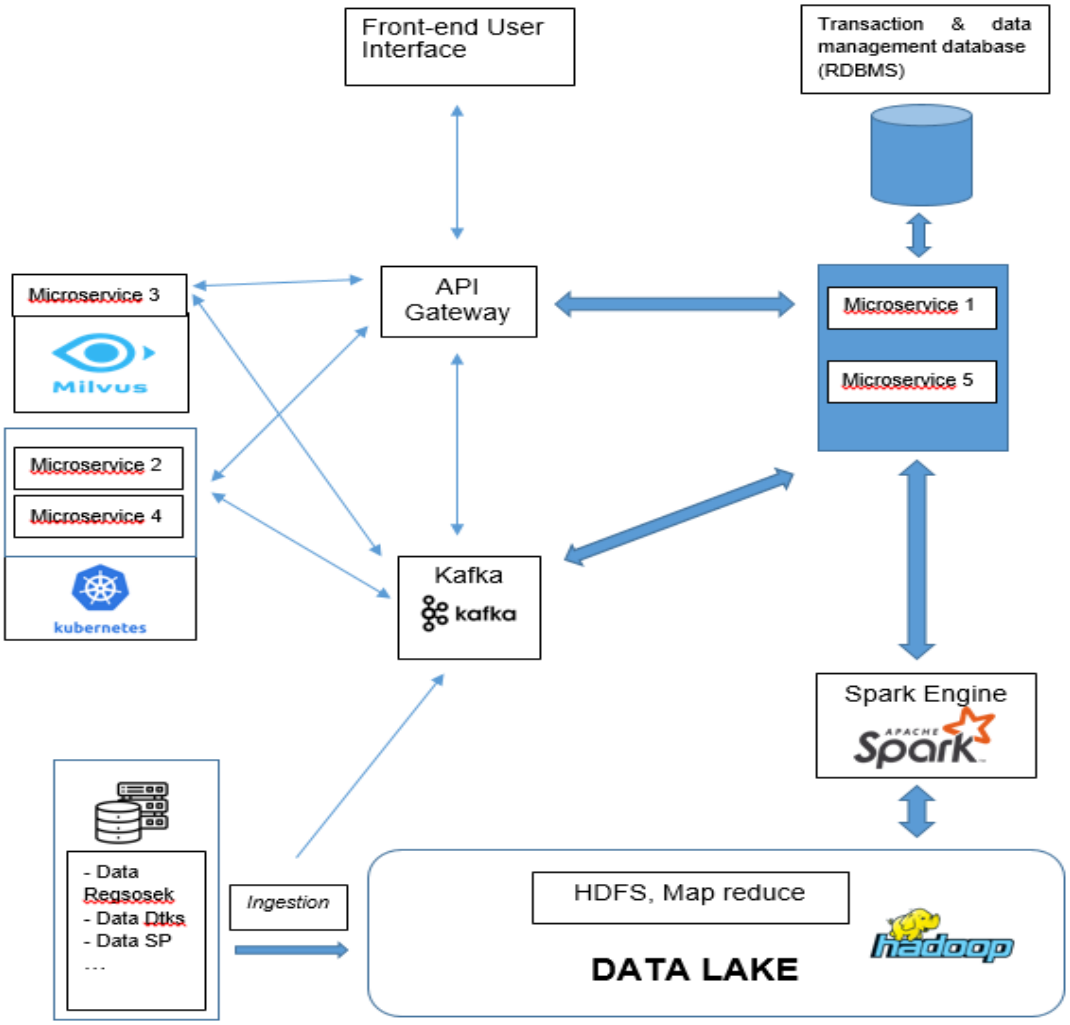
5. Selanjutnya algoritma pencocokan entitas dengan menggunakan metode Ditto diterapkan pada kandidat pasangan untuk mengidentifikasi dan mengklasifikasikan pasangan entitas mana yang benar-benar match dan mewakili entitas yang sama di dunia nyata. Proses ini akan berakhir dengan memberikan probabilitas kecocokan untuk masing-masing kandidat yang diklasifikasikan sebagai cocok/match.

---

[2] Li, Yuliang, et al. "Deep entity matching with pre-trained language models." *arXiv preprint arXiv:2004.00584* (2020). https://arxiv.org/pdf/2004.00584.pdf

Arsitektur Sistem Aplikasi

Secara umum, arsitektur aplikasi yang diusulkan terdiri dari lima komponen, yaitu *frontend user interface*, *microservices*, database transaksi dan manajemen data, *streaming engine*, dan *data lake*. *Frontend user interface* menggunakan javascript dengan framework solid.js, yang mendukung penyediaan user interface yang dinamis dengan performa yang baik. Di sisi lain, frontend application ini dilayani oleh *microservices* yang dibangun menggunakan java dan python, serta menggunakan *vector database* milvus, untuk mendukung penyimpanan, pemrosesan dan retrieval *vector data*, yang bersesuaian dengan metode dan konsep yang digunakan.

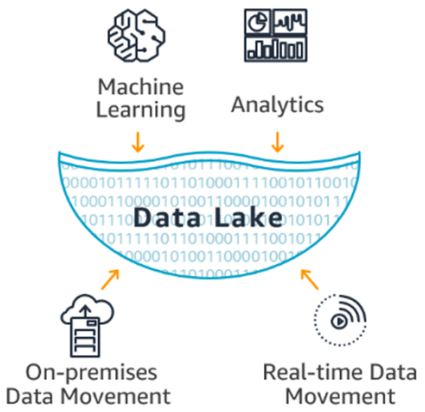


Gambar 2. Arsitektur aplikasi

Untuk mendukung berjalannya aplikasi, diperlukan pula sebuah RDBMS (*Relational Database Management System*) untuk menjamin semua transaksi yang dilakukan pada aplikasi dikerjakan dengan akurat dan konsisten. Database ini digunakan sebagai acuan dan gambaran *current state* dari aplikasi dan proses-proses yang sedang berjalan pada aplikasi. Selanjutnya, ada streaming engine, berupa Kafka sebagai event streaming engine, yang dapat men *trigger service-service* pada aplikasi, agar dapat berjalan pada saat tertentu yang telah ditentukan, dan juga memberikan pesan pada sistem saat proses tersebut telah selesai. Selain itu, streaming engine lainnya adalah Spark, sebagai

real time data processing dan *streaming engine* yang dapat melakukan proses penyimpanan, *retrieval* dan *processing* data dari dan ke *data lake*. Komponen terakhir adalah *data lake*, yaitu sebuah sistem penyimpanan data umum yang dapat digunakan untuk menyimpan berbagai macam jenis data dalam ukuran besar secara terklusterisasi. Pada aplikasi ini, *data lake* digunakan sebagai penyimpanan data dari berbagai sumber, yang telah dilakukan *ingestion*. Selanjutnya data-data tersebut akan di *retrieve* dan diproses oleh *service-service* yang tersedia pada aplikasi. Layanan data lake yang digunakan berupa HDFS (*hadoop distributed file system*) dan *Map reduce* yang telah tersedia pada apache hadoop.

Aplikasi dan layanan yang diusulkan ini akan berjalan di atas sebuah sistem *data lake*, dimana data yang digunakan telah terkumpul dan dapat diakses oleh aplikasi pada sistem *data lake* tersebut. Data keluaran dari aplikasi, berupa data yang telah selesai di deduplikasi maupun disambungkan dengan data-data lain, akan dimasukan kembali ke sistem *data lake* tersebut untuk dapat dilakukan *retrieval* kembali, baik oleh aplikasi ini maupun aplikasi SEPAKAT dan aplikasi lainnya yang berkaitan. Deskripsi dari arsitektur *data lake* dapat digambarkan dari arsitektur *datalake* yang ditawarkan oleh AWS Amazon Cloud[3].



Gambar 3. Gambaran arsitektur *data lake*

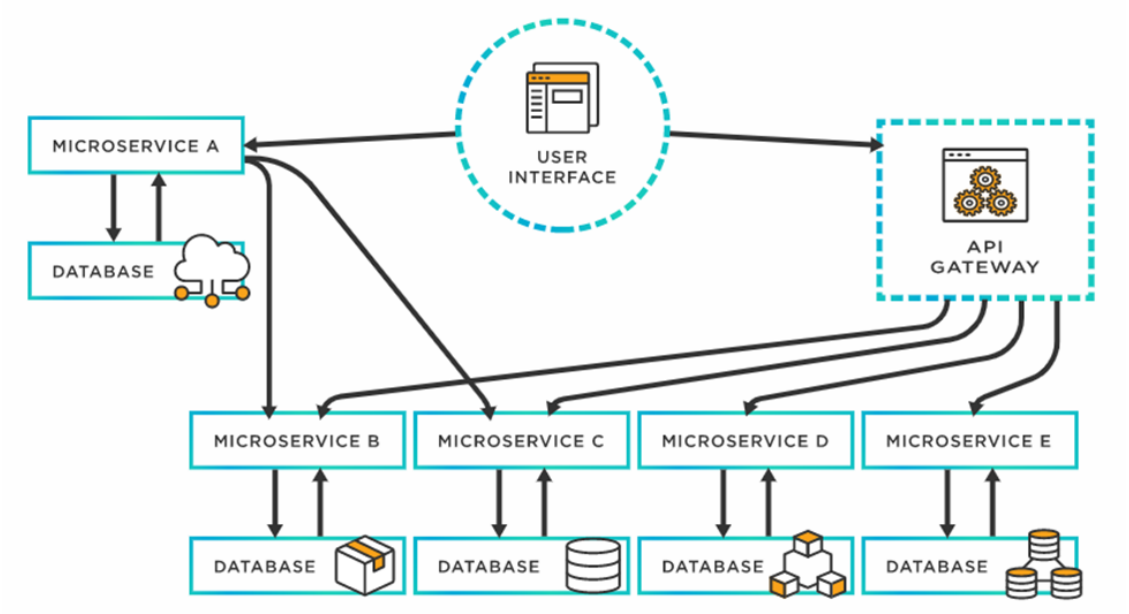
Modul integrasi ini dirancang menggunakan arsitektur *microservices* sebagai basis layanan utamanya. Selain itu terdapat *API gateway* yang berguna untuk menyambungkan aplikasi *frontend/UI* ke servis-servis yang tersedia. Arsitektur microservices memiliki keunggulan antara lain:

- Meningkatkan skalabilitas dan fleksibilitas, karena setiap *microservice* dapat dikembangkan dan diterapkan secara independen.
- Memudahkan pemeliharaan dan perawatan, karena setiap *microservice* dapat diperbaharui atau dimodifikasi tanpa mempengaruhi keseluruhan sistem
- Meningkatkan efisiensi dan kecepatan pengembangan, karena setiap tim pengembang dapat fokus pada pengembangan *microservice* tertentu.

[3] https://aws.amazon.com/id/big-data/what-is-a-data-lake/

- Memudahkan pemeliharaan dan perawatan, karena setiap *microservice* dapat dijalankan pada lingkungan yang berbeda.
- Meningkatkan keamanan, karena setiap *microservice* dapat dikelola secara terpisah dan dapat diberikan autentikasi dan otorisasi yang berbeda.

Berikut adalah gambaran arsitektur microservices disadur dari Tibco[4].



Gambar 4. Gambaran arsitektur *microservices*

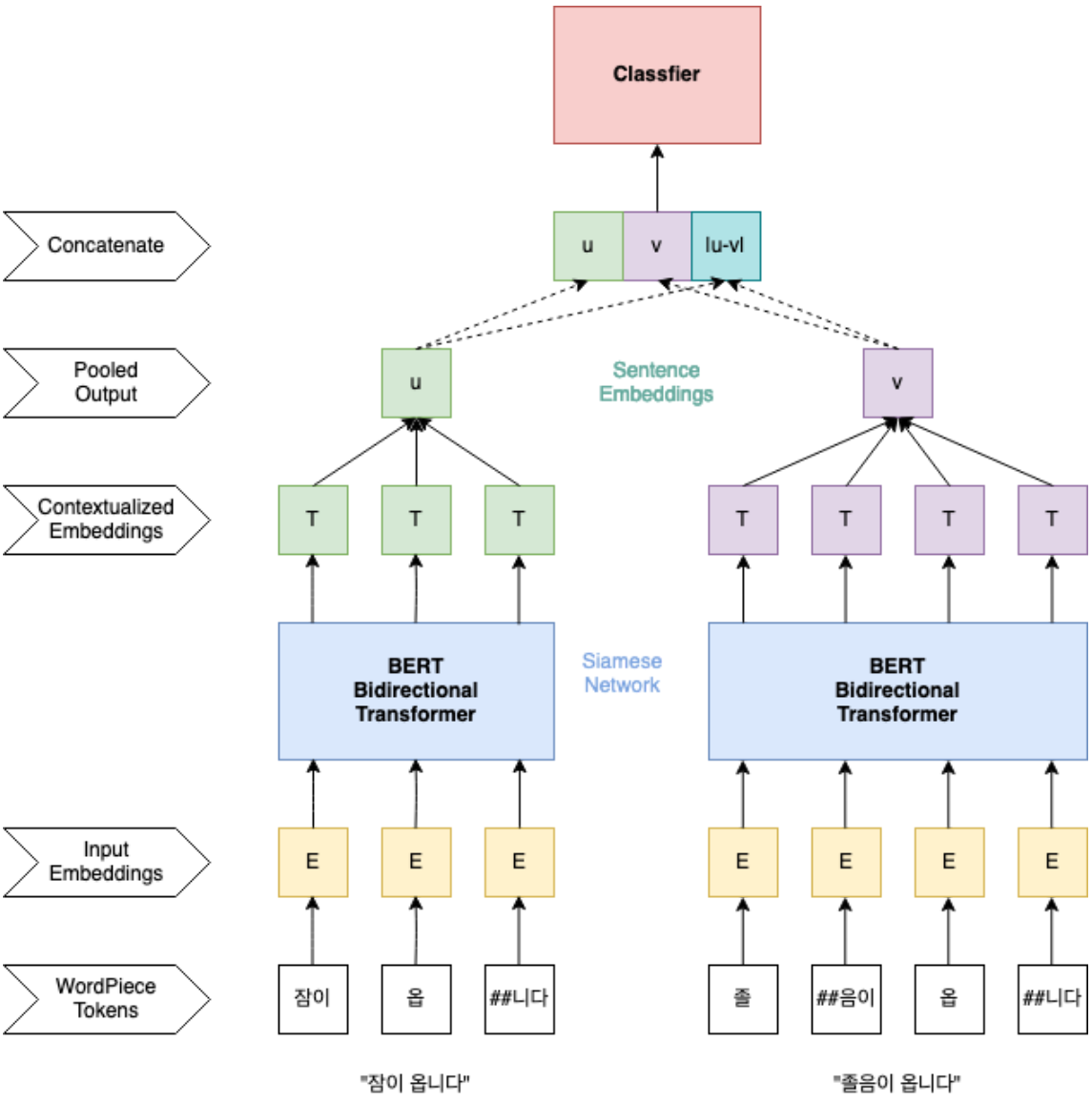
Arsitektur sistem resolusi entitas umumnya terdiri dari beberapa komponen utama, termasuk sumber data, modul *entity extraction*, algoritma *entity matching*, dan modul output hasil. Berikut adalah uraian singkat untuk masing-masing layanan:

1. *Microservice* pertama dalam sistem ini adalah layanan pengumpulan dan pemrosesan data. Layanan ini bertanggung jawab untuk mengumpulkan dan membersihkan data dari berbagai sumber, seperti database, web API, dan dokumen teks. Ini memastikan bahwa data dalam format yang konsisten dan siap untuk resolusi entitas. Untuk mengolah data, setiap entitas akan dijadikan ancaman sebagai kalimat dengan menggabungkan nilai atribut dari setiap entitas, seperti nama, umur, dan jenis kelamin ke dalam satu rangkaian kata.

2. *Microservice* kedua adalah prosedur *entity extraction* dan *embedding generation*. Layanan ini menggunakan algoritma Natural Language Processing (NLP) untuk mengekstrak entitas dari data dan mengubahnya menjadi *sentence embedding* menggunakan library pada bahasa Python, yaitu *SentenceBERT* dengan menggunakan IndoBERT hasil studi dari Wilie dkk. (2020)[5] sebagai model dasar. *SentenceBERT* menerapkan arsitektur *siamese network* untuk melatih model deep

[4] https://www.tibco.com/reference-center/what-are-microservices
[5] Wilie, Bryan et. al. "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding" arXiv:2009.05387 (2020). https://arxiv.org/pdf/2009.05387.pdf

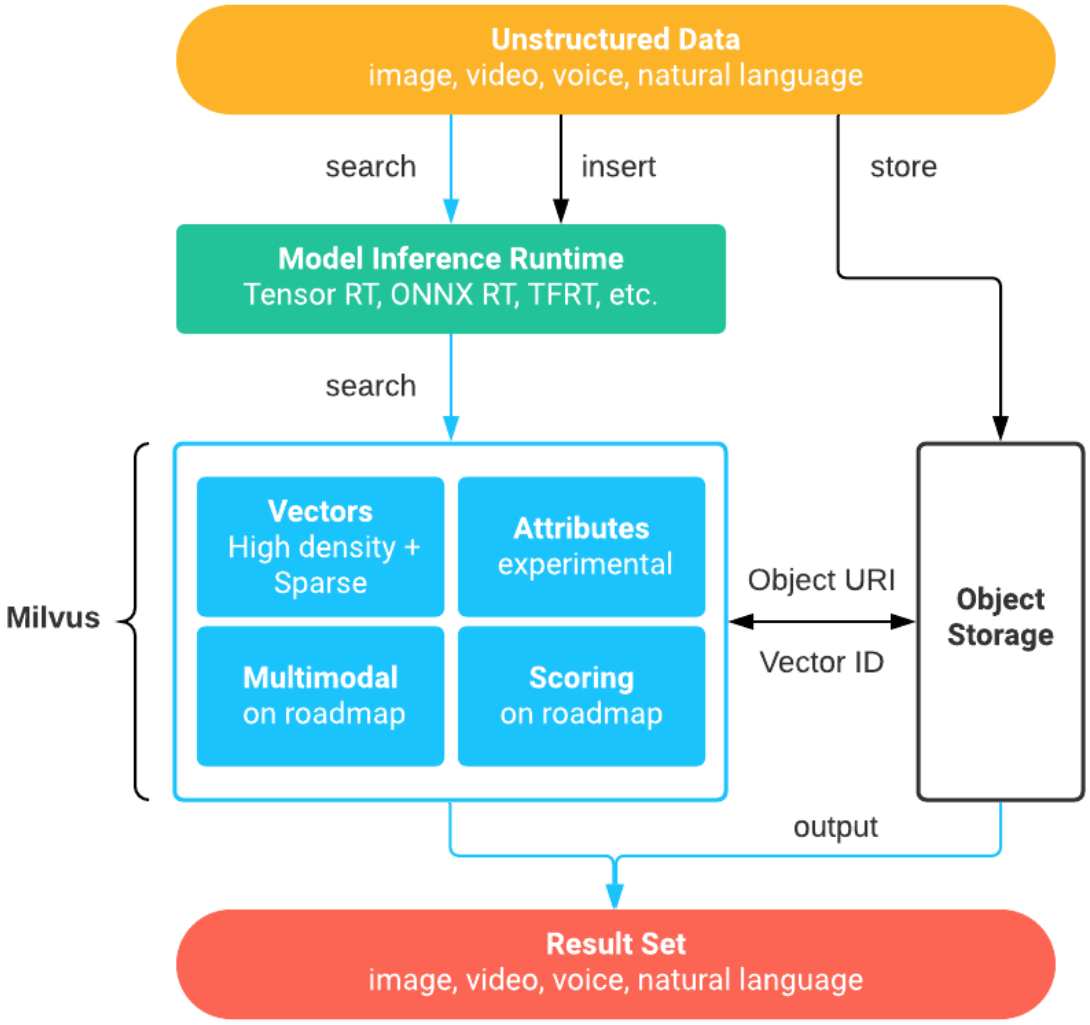
learning. *Sentence embedding* adalah proses merepresentasikan kalimat dalam bentuk vektor numerik, yang menangkap makna semantik dan konteks yang terdapat pada data berupa teks. Hal Ini membuat operasi matematika yang efisien pada representasi kalimat tersebut dapat dilakukan, seperti melakukan analisis perbandingan atau pengelompokan/clustering, dan dapat diaplikasikan di berbagai kasus NLP seperti analisis sentimen maupun kemiripan teks. Arsitektur *SentenceBERT* disadur dari @snulp[6] yang telah mengembangkannya untuk bahasa Korea.



Gambar 4. Arsitektur SentenceBERT untuk bahasa Korea

[6] https://github.com/snunlp/KR-SBERT

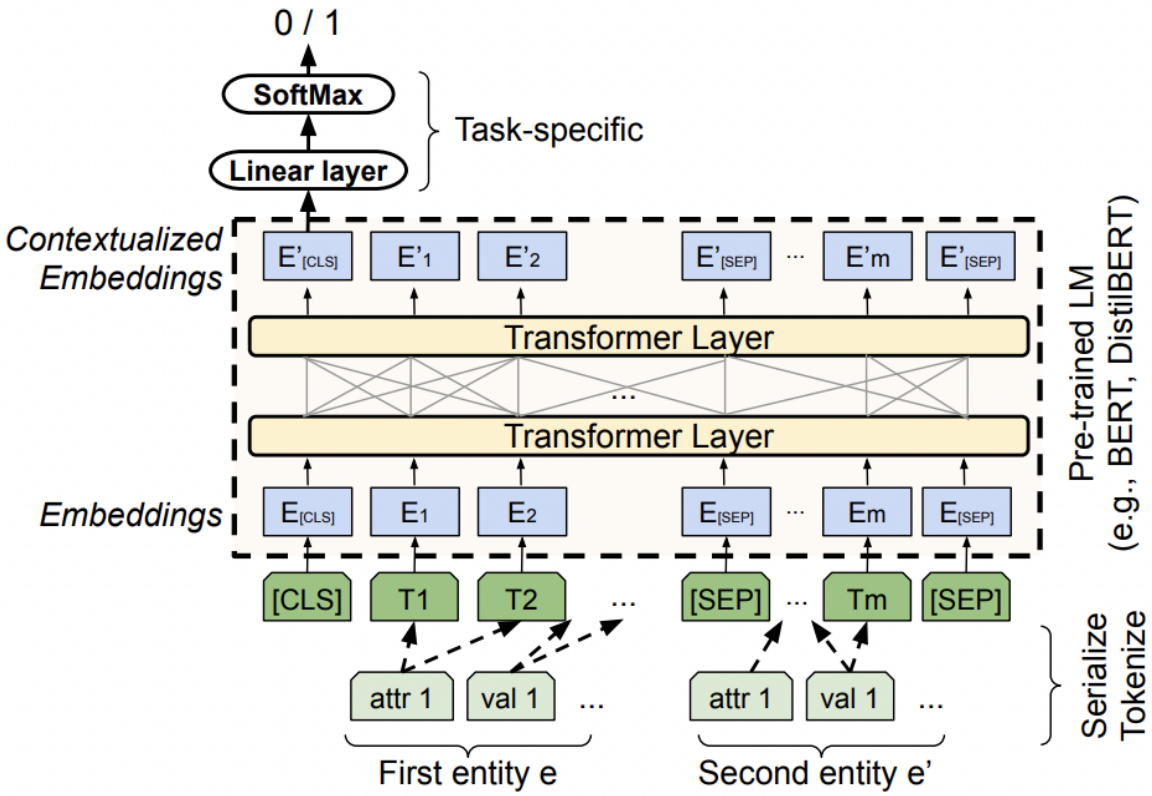
3. *Microservice* berikutnya adalah *vector database*. Berbeda dengan database biasa, *vector database* dirancang khusus untuk pemanfaatan data berbentuk *vector* sebagaimana wujud representasi dari *sentence embedding*. *Sentence embedding* tersebut akan disimpan ke dalam database vektor menggunakan database engine yaitu *Milvus vector database*.
   Milvus adalah database *open-source* yang memungkinkan pengguna untuk menyimpan dan melakukan pencarian vektor dalam skala besar untuk berbagai aplikasi seperti sistem rekomendasi dan mesin pencari, termasuk untuk *sentence similarity*. Teknologi ini dapat menggunakan penyimpanan terdistribusi (*distributed storage*) dan menerapkan *cloud architecture* untuk mengaktifkan skalabilitas dan kinerja tinggi, dan menawarkan fitur seperti pengindeksan vektor, pencarian kesamaan (similarity), dan partisi data multidimensi. Arsitektur umum database vektor disadur dari dokumentasi Milvus[7].



Gambar 5. Arsitektur Umum Milvus

[7] https://milvus.io/docs/v1.1.1/overview.md

4. *Microservice* keempat adalah layanan pencocokan entitas (*entity matching*) dan deduplikasi. Layanan ini akan menggunakan metode deep learning, yaitu *Ditto*: *Deep Entity Matching with Pre-Trained Language Models* yang merupakan hasil studi dari Li, dkk. pada tahun 2020. Kami akan menggunakan *IndoBERT* sebagai pre-trained language modelnya karena *IndoBERT* telah dilatih menggunakan bahasa Indonesia dari dataset wikipedia, sehingga mampu memahami arti semantik dari kata-kata yang ada dalam bahasa Indonesia. Arsitektur disadur dari Li, dkk. (2020)[8]



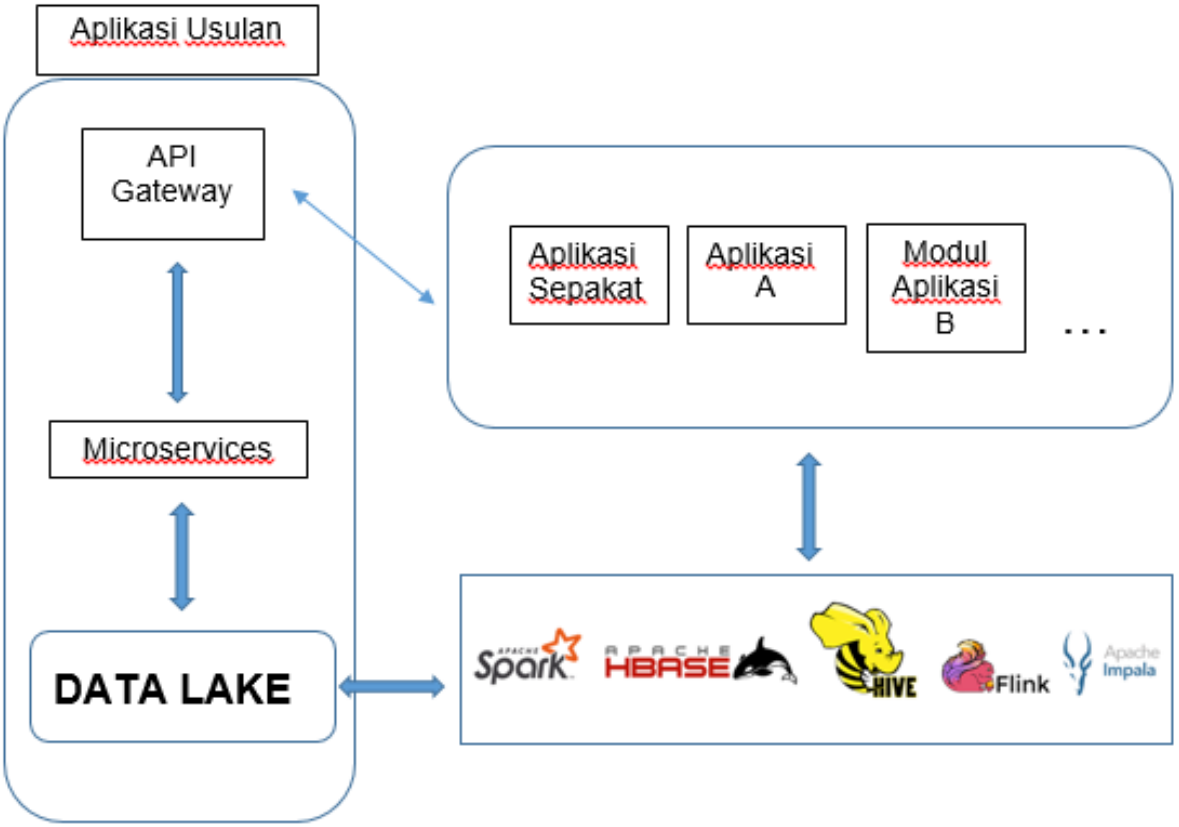
Gambar 6. Arsitektur model Ditto

5. *Microservice* terakhir adalah interface dari sistem resolusi entitas dan layanan *post-processing data* berbasis aplikasi web. Layanan ini akan melakukan finalisasi entitas dan menggabungkannya. Layanan ini juga melakukan *post-processing* yang diperlukan, seperti membuat dashboard, laporan, dan memperbaharui basis data. Pengguna juga dapat berinteraksi dengan sistem untuk melakukan pengecekan hasil dan kualitas integrasi data.

Aplikasi yang diusulkan, terbuka dan siap untuk diintegrasikan dengan aplikasi lain. Arsitekturnya yang berbasis *microservices* membuatnya mudah untuk berkomunikasi dan saling berbagi pakai dengan modul ataupun sistem yang lain. Dalam hal ini, dicontohkan aplikasi Sepakat, yang menyediakan fitur analisis data dan dashboard untuk menunjang pengentasan kemiskinan di indonesia. Dalam hal ini aplikasi sepakat dapat mendapatkan data Regsosek atau data-data lain yang

[8] Li, Yuliang, et al. "Deep entity matching with pre-trained language models." *arXiv preprint arXiv:2004.00584* (2020). https://arxiv.org/pdf/2004.00584.pdf

telah dilakukan deduplikasi dan memiliki tautan (linkage) dengan data lain, misalkan data DTKS, data Kartu Pra-kerja, Data PLN, Data BPJS, Data My Pertamina, dll. Layanan serta data tersebut dapat didapatkan dengan berkomunikasi pada sistem aplikasi yang diusulkan.

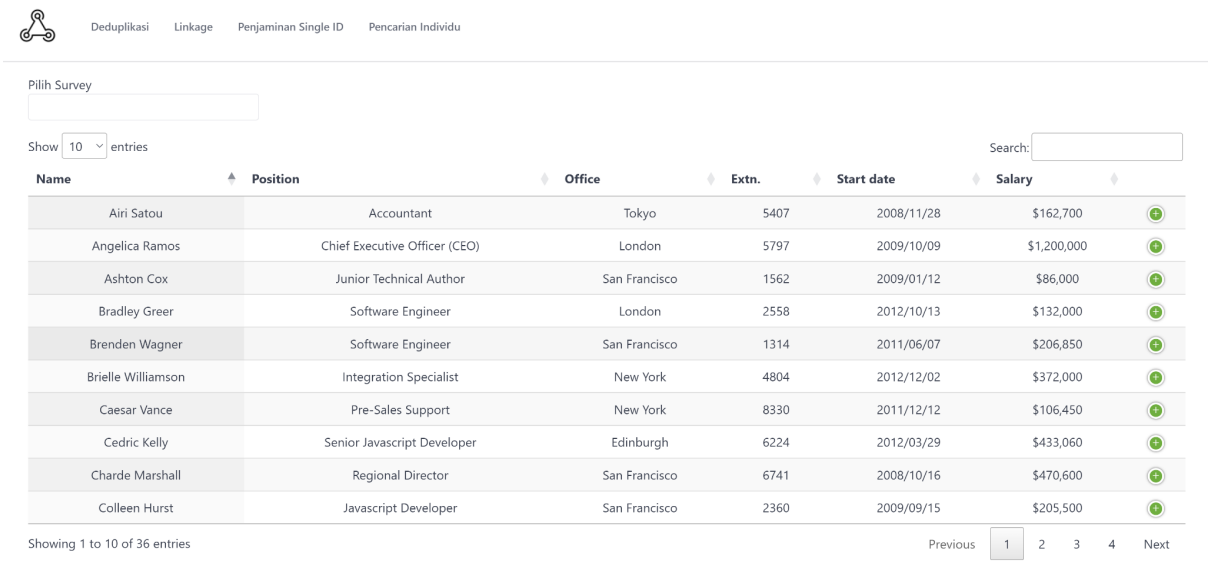


Gambar 7. Arsitektur integrasi antar aplikasi

Selain itu, *data lake* yang menjadi basis penyimpanan data dari sistem aplikasi ini memiliki kapasitas untuk diperbesar secara horizontal (*horizontal scale-up*) melalui penambahan worker nodes yang tergabung dalam *cluster computing*, sehingga sangat tepat digunakan untuk manajemen data nasional dan sangat mendukung untuk digunakan sebagai wadah saling bagi pakai data antar aplikasi, modul ataupun sistem yang lain dengan menggunakan berbagai macam *Big Data tools*. Sehingga, manfaat dari output data aplikasi yang dihasilkan, dapat lebih luas dan terus berkembang.

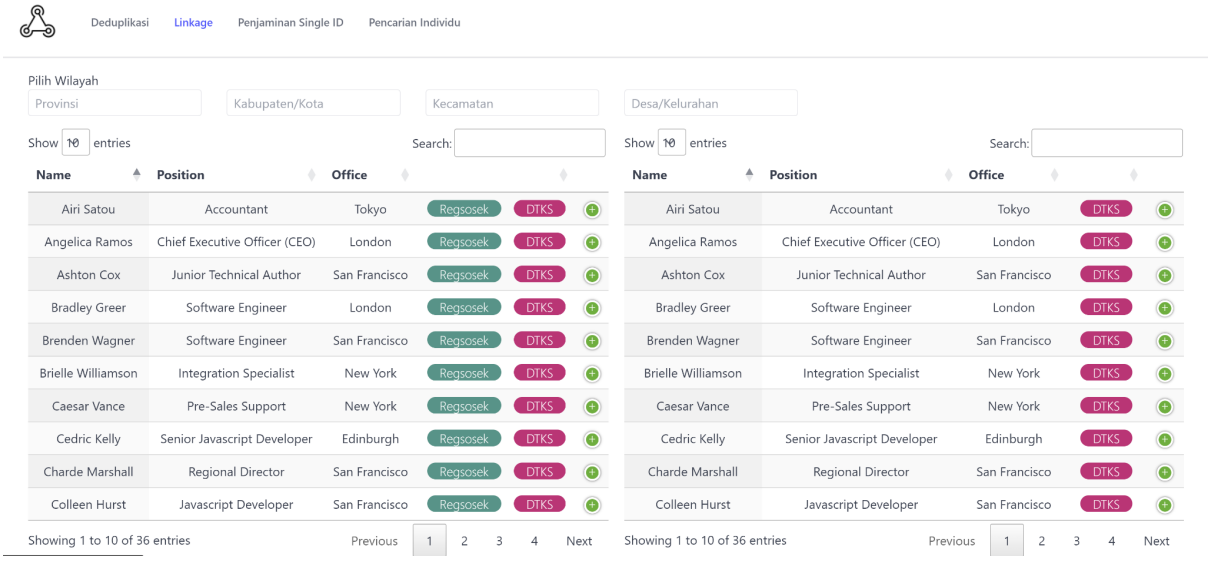
Rancangan *User Interface*

Deduplikasi



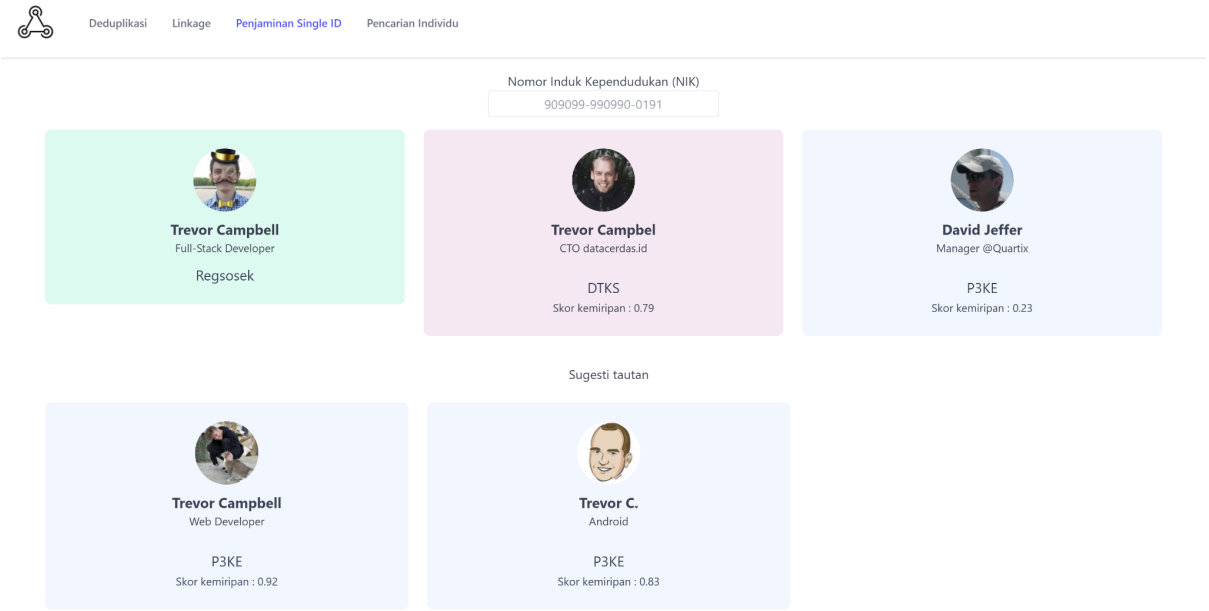
Pada menu ini akan menampilkan potensi duplikasi record untuk setiap data yang dipilih. Aplikasi selanjutnya akan menampilkan daftar record yang teridentifikasi berpotensi duplikat. Adapun pengguna juga diijinkan untuk melihat rincian lebih detail masing-masing individu tersebut. Selanjutnya, pengguna diberikan opsi untuk melakukan konfirmasi atas hasil temuan tersebut.

Linkage



Menu ini akan menampilkan record data menurut wilayah yang dipilih user serta tautan (linkage) padanan individu pada data sumber data lain yang tersedia.

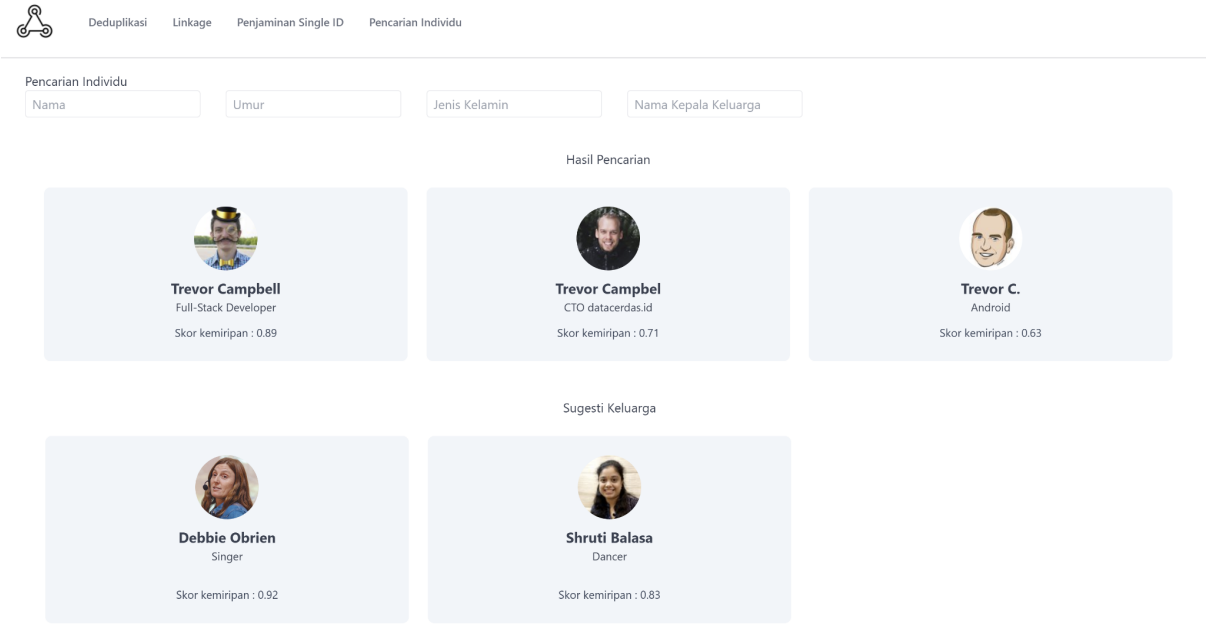
Penjaminan Single ID



Menu ini menyediakan fitur pencarian berdasarkan NIK, diikuti tampilan record yang dianggap memiliki tautan (*linkage*) terhadap NIK tersebut. Selain, turut menampilkan skor *linkage* record tersebut. Skor yang tinggi dianggap bahwa konsistensi NIK tersebut sudah cukup tinggi pada data-data yang tersedia. Apabila skor *linkage* rendah, aplikasi akan menampilkan hasil pentautan (*linkage*) oleh sistem yang memiliki skor terbaik untuk rekomendasi pasangan data tersebut.

Misalkan pada tampilan di atas, disajikan data individu berdasarkan NIK, dari data Regsosek, DTKS, dan P3KE. Dari data yang tersedia tersebut, diketahui bahwa skor kemiripan antara data Regsosek sebagai data dasar, dengan data DTKS sebesar 0.79, hal ini menunjukan bahwa data tersebut telah konsisten, namun skor kemiripan dengan P3KE rendah, yaitu 0.23, hal ini menunjukan bahwa sebenarnya tautan (*linkage*) antara data regsosek dan P3KE dengan dasar NIK tersebut rendah, artinya berpotensi untuk salah. Selanjutnya, aplikasi akan menampilkan rekomendasi yang sesuai, berdasarkan penghitungan *linkage* oleh sistem menggunakan metode yang telah dijelaskan di atas.

Pencarian Individu



Menu ini menyediakan fitur bagi pengguna untuk melakukan pencarian secara realtime dari data individu dengan menginputkan karakteristik dari individu yang dicari, baik itu NIK sebagai ID maupun karakteristik lain seperti Nama, Umur, Jenis kelamin, dan Nama Kepala Keluarga. Selanjutnya aplikasi akan menampilkan hasil pencarian tersebut, termasuk hasil integrasi pada sumber data lain.

Fitur ini bisa dianalogikan secara lebih sederhana seolah-olah kita sedang melakukan pencarian produk di suatu *platform e-commerce* dimana individu dapat dianalogikan sebagai suatu produk. Dengan memasukkan kata kunci tertentu, maka aplikasi akan memberikan daftar produk yang serupa/mirip dari berbagai pedagang online dengan kata kunci yang diinginkan, dimana pedagang dalam hal ini kurang lebih sama dengan sumber datanya.

Timeline Pengembangan Sistem

Berdasarkan rancangan arsitektur yang telah ditentukan, sistem aplikasi ini direncanakan untuk dapat diselesaikan dalam periode waktu 20 minggu, atau sekitar 5-6 bulan dengan rincian sebagai berikut:

Kegiatan	Minggu																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Persiapan	■	■	■	■																
Identifikasi Data Pendukung	■	■																		
Pengumpulan Studi Literatur	■	■	■	■																
Identifikasi Infrastruktur		■	■	■																
Investigasi *Platform* Existing			■	■																
Analisis			■	■	■	■	■	■												
Pengumpulan Kebutuhan			■	■	■	■	■	■												
Perumusan Arsitektur			■	■	■	■	■	■												
Kajian Model pada Data Pendukung					■	■	■	■												
Penentuan Konsep Solusi					■	■	■	■												
Analisis Librari Pendukung							■	■												
Perancangan								■	■	■	■	■	■	■						
Perancangan Solusi Sistem								■	■	■	■									
Perancangan Tampilan Antarmuka									■	■	■	■								
Perancangan Database Transaksi									■	■	■	■								
Perancangan Data Lake									■	■	■	■	■	■						
Pereancangan Database Vektor									■	■	■	■	■	■						
Perancangan Model									■	■	■	■	■	■						
Implementasi									■	■	■	■	■	■	■	■	■	■	■	■
Pembangunan									■	■	■	■	■	■	■	■	■	■	■	■
Testing															■	■	■	■	■	■
Dokumentasi dan Deploy																	■	■	■	■

Prospek Hasil Implementasi

- Menghasilkan basis data individu nasional yang bersih dari duplikasi.
- Menghasilkan record data yang dapat tersambungkan/ditautkan dengan record data-data lainnya, sehingga bersifat komprehensif.
- Mendukung penyediaan data individu berbasis Single ID number (NIK).
- Menyediakan sistem aplikasi yang terbuka dan siap diintegrasikan dengan aplikasi, modul, atau sistem lainnya.

Contoh hasil pentautan data menggunakan data contoh adalah sebagai berikut:

Tabel 1. Contoh Hasil Pentautan

NIK A	Nama A	Nama KRT A	Jenis Kelamin A	Umur A	NIK B	Nama B	Nama KRT B	Jenis Kelamin B	Umur B	Kemiripan (0..1)	Flag
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
990318260***	m humaidi hamzani	mustapa	L	1	99999999***	m humaidi hamzani	mustapa	L	1	0.98	Match
990318701***	salminah	salminah	P	72	99031870***	salminah	salminah	P	61	0.91	Match
990318020***	umar	salminah	L	23	99031802***	umar	salminah	L	24	0.90	Match
990318111***	abidin ali	abidin ali	L	36	99031802***	ahmad yusup	ahmad yusup	L	36	0.12	Not Match

Penjelasan:

Pada tabel diatas disajikan hasil pentautan antara sumber data A dan sumber data B. Diperoleh tiga pasangan match dan 1 pasangan tidak match. Contoh pasangan match adalah salminah dengan skor kemiripan berdasarkan prediksi machine learning sebesar 0.91. Hal ini didasarkan karena nama, nama KRT, dan jenis kelaminnya sama di kedua data meskipun umurnya sedikit berbeda namun masih dalam rentang toleransi. Sedangkan untuk contoh pasangan tidak match adalah antara abidin ali dan ahmad yusup dengan skor kemiripan 0.12 karena hanya umur dan jenis kelaminnya yang sama, sedangkan namanya benar-benar berbeda.

Potensi Pengembangan

- Penambahan karakteristik dalam pembentukan embedding/vector representation setiap individu dengan menggunakan bio-characteristic seperti fingerprint, retina, atau face recognition sehingga menghasilkan integrasi data yang lebih berkualitas.
- *Linkage* data dapat dikembangkan dan dilakukan secara lintas waktu, contohnya untuk data regsosek 2022 dengan data Sensus Penduduk 2020 ataupun data Sensus Penduduk 2010

yang diintegrasikan dengan data geospasial/kewilayahan untuk melihat dinamika migrasi penduduk dan proyeksi jumlah penduduk.

- *Linkage* data dapat dilakukan secara berkesinambungan dalam skala besar dengan berbagai sumber data lainnya termasuk dari sektor privat, sesuai dengan rancangan arsitektur sistem aplikasi integrasi data yang dirancang secara terklusterisasi sehingga memiliki kapasitas upscaling secara horizontal.
- *Linkage* di level keluarga/rumah tangga dapat mendeteksi dan mengantisipasi pemberian bantuan atau program yang kurang tepat sasaran/redundan di dalam lingkup keluarga/rumah tangga.
- Mendukung penyediaan satu data Indonesia yang berkualitas dan one single ID yang bebas redundansi melalui mekanisme deduplikasi menggunakan metode probabilistic *linkage*.
- Terbukanya peluang analisis dan estimasi data berdasarkan data terintegrasi.
- Integrasi data dapat dilakukan antara data dasar Regsosek dengan data berbasis survei sampling sehingga dapat digunakan untuk meningkatkan kualitas estimasi survei menggunakan metode *post calibration*.

Batasan

- Belum ada ketersediaan data yang real untuk dilakukan analisis dan experiment skala besar.
- Kualitas integrasi data bergantung pada ketersediaan karakteristik dari setiap data beserta tingkat ketersediaannya.

Daftar Pustaka

Zhu, Ying, et al. "When to conduct probabilistic linkage vs. deterministic linkage? A simulation study." *Journal of biomedical informatics 56* (2015): 80-86. https://doi.org/10.1016/j.jbi.2015.05.012

Li, Yuliang, et al. "Deep entity matching with pre-trained language models." *arXiv preprint arXiv:2004.00584 (2020).* https://arxiv.org/pdf/2004.00584.pdf

amazon.com. "What is a data lake?" Diakses pada tanggal 13 Desember 2022, dari https://aws.amazon.com/id/big-data/what-is-a-data-lake/

tibco.com. "What are Microservices?" Diakses pada tanggal 13 Desember 2022, dari https://www.tibco.com/reference-center/what-are-microservices

Wilie, Bryan et. al. "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding" arXiv:2009.05387 (2020). https://arxiv.org/pdf/2009.05387.pdf

Park, Suzi and Hyopil Shin. "KR-SBERT: A Pre-trained Korean-specific Sentence-BERT model". Github (2021). Diakses pada tanggal 14 Desember 2022, dari https://github.com/snunlp/KR-SBERT

milvus.io. "What is milvus?" Diakses pada tanggal 14 Desember 2022, dari https://milvus.io/docs/v1.1.1/overview.md