Bab 4
ANALITIK & LIFECYCLE BIG DATA

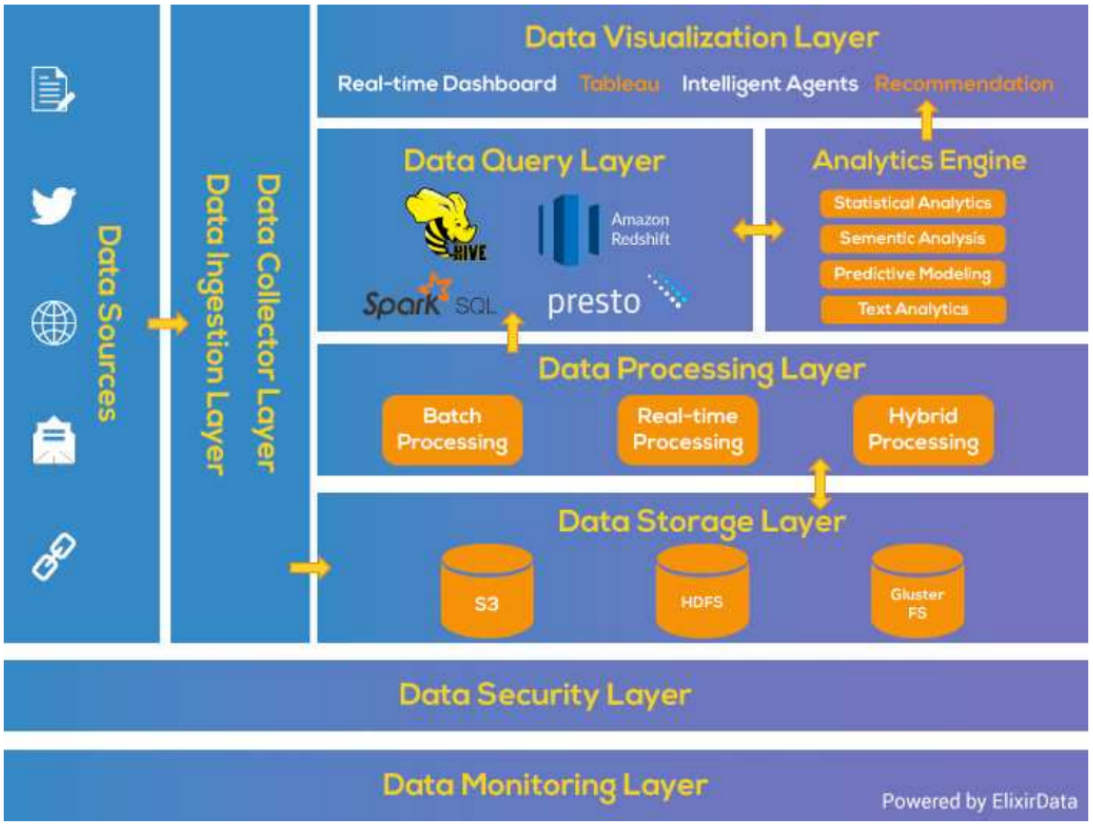
Dani Anggoro, S.Kom., M.Kom

4.1 Pengertian

Analitik dan *lifecycle* Big Data merujuk pada serangkaian langkah atau tahapan yang ditempuh oleh data sepanjang perjalanannya dari pengumpulan hingga analisis dan penggunaan akhir. Informasi apa saja yang bisa digali dari Big Data pada perusahaan dan strategi apa saja yang bisa dilakukan dari masing-masing perusahaan adalah bagian dari analitik dan *lifecycle* Big Data

4.2 Arsitektur Big Data

Arsitektur big data adalah pada proses pengumpulan, penyimpanan, pengolahan, analisis, dan pengelolaan data berskala besar. Cara Terbaik untuk mendapatkan solusi dari Permasalahan Big Data (Big Data Solution) adalah dengan "Membagi Masalahnya". Big Data Solution dapat dipahami dengan baik menggunakan Layered Architecture. Arsitektur Layered dibagi ke dalam Lapisan yang berbeda dimana setiap lapisan memiliki spesifikasi dalam melakukan fungsi tertentu. Arsitektur tersebut membantu dalam merancang Data Pipeline (jalur data) dengan berbagai mode, baik Batch Processing System atau Stream Processing System..



Gambar 4.1 Arsitektur Big Data

Arsitektur ini terdiri dari 6 lapisan yang menjamin arus data yang optimal dan aman

A. Data Ingestion Layer

Lapisan ini merupakan langkah awal untuk data yang berasal dari sumber tertentu dan akan memulai perjalanannya. Data disini akan dilakukan pemrioritasan dan pengkategorian, sehingga data dapat diproses dengan mudah diteruskan ke lapisan lebih lanjut.

Tool yang dapat digunakan, yaitu Apache Flume, Apache Nifi (Pengumpulan dan Penggalian Data dari Twitter menggunakan Apache NiFi untuk Membangun Data Lake),

Data masuk ke dalam Data Lake dalam bentuk mentah, dan semua datanya disimpan, tidak hanya data yang digunakan saja, tapi juga data yang mungkin digunakan di masa depan. Di Data Lake semua data tersimpan dalam bentuk aslinya.

B. Data Collector Layer

Di Lapisan ini, lebih fokus pada penyaluran data dari lapisan penyerapan atau pengambilan data awal (ingestion) ke jalur data yang lainnya. Pada Lapisan ini, data akan dipisahkan sesuai dengan kelompoknya atau komponen-komponennya (Topik: kategori yang ditentukan pengguna yang pesannya dipublikasikan, Produser - Produsen yang memposting pesan dalam satu topik atau lebih, Konsumen - berlangganan topik dan memproses pesan yang diposkan, Brokers - Pialang yang tekun dalam mengelola dan replikasi data pesan) sehingga proses analitik bisa dimulai.

Tool yang dapat digunakan, yaitu Apache Kafka.

C. Data Processing Layer

Fokus utama lapisan ini adalah untuk sistem pemrosesan data pipeline atau dapat kita katakan bahwa data yang telah kita kumpulkan di lapisan sebelumnya akan diproses di lapisan ini. Di sini kita melakukan ekstraksi dan juga learning dari data untuk diarahkan ke tujuan yang bermacam-macam, mengklasifikasikan arus data yang seharusnya diarahkan dan ini adalah titik awal di mana analitik telah dilakukan.Data pipeline merupakan komponen utama dari Integrasi Data. Data pipeline mengalirkan dan mengubah data real-time ke layanan yang memerlukannya, mengotomatiskan pergerakan dan transformasi data, mengolah data yang berjalan di dalam aplikasi Anda, dan mentransformasikan semua data yang masuk ke dalam format standar sehingga bisa digunakan untuk analisis dan visualisasi. Jadi, Data pipeline adalah

rangkaian langkah yang ditempuh oleh data Anda. Output dari satu langkah dalam proses menjadi input berikutnya. Langkah-langkah dari Data pipeline dapat mencakup pembersihan, transformasi, penggabungan, pemodelan dan banyak lagi, dalam bentuk kombinasi apapun.

Tool yang dapat digunakan, yaitu Apache Sqoop, Apache Storm, Apache Spark, Apache Flink.

D. Data Storage Layer

Media penyimpanan menjadi tantangan utama, saat ukuran data yang digunakan menjadi sangat besar. Lapisan ini berfokus pada "tempat menyimpan data yang begitu besar secara efisien".

Tool yang dapat digunakan, yaitu Apache Hadoop (HDFS), Gluster file systems (GFS), Amazon S3.

E. Data Query Layer

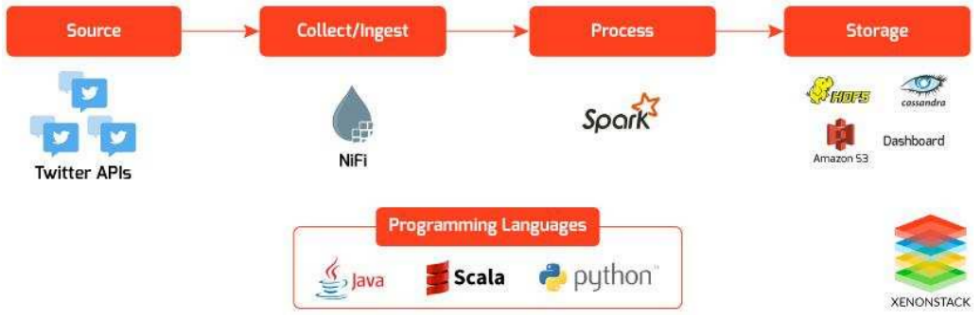
Lapisan ini merupakan tempat berlangsungnya pemrosesan secara analitik yang sedang dalam keadaan aktif. Di sini, fokus utamanya adalah mengumpulkan data value sehingga dapat dibuat lebih bermanfaat dan mudah digunakan untuk lapisan berikutnya.

Tool yang dapat digunakan, yaitu Apache Hive, Apache (Spark SQL), Amazon Redshift, Presto.

F. Data Visualization Layer

Proses Visualisasi, atau tahapan merepresentasikan data dalam bentuk visual, kemungkinan ini adalah tingkat yang paling bergengsi, di mana pengguna data pipeline dapat merasakan hasil laporan yang mendetail dan mudah dipahami dari data value yang telah disosialisasikan. Kita membutuhkan sesuatu yang

akan menarik perhatian orang dari visualisasi data, sehingga membuat temuan Anda mudah dipahami dengan baik oleh mereka melalui visualisasi tersebut.



Gambar 4.2 Integrating Apache Spark dan NiFi for Data Lakes

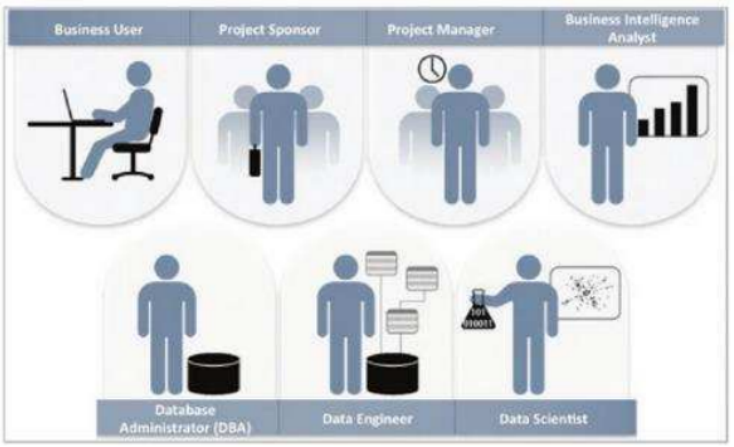
Tool yang dapat digunakan, yaitu Tableau, Kibana sebagai Real-Time Dashboards, Angular.js sed'agai Intelligence Agents misalnya agen dapat mengingat hal-hal yang mungkin Anda sudah lupa, dengan cerdas meringkas data yang kompleks, belajar dari Anda dan bahkan membuat rekomendasi untuk Anda, menemukan dan memfilter informasi saat Anda melihat data perusahaan atau berselancar di Internet dan tidak tahu di mana informasi yang tepat, React.js sed'agai sistem recommender untuk memprediksi tentang kriteria pengguna, yaitu menentukan model penggunanya seperti apa.

Apache Spark digunakan secara luas untuk pengolahan Big Data. Spark bisa mengolah data di dua mode yaitu Pengolahan Batch Mode dan Streaming Mode. Apache NiFi ke Apache Spark melakukan transmisi data menggunakan komunikasi situs ke situs. Dan output port-nya digunakan untuk mempublikasikan data dari sumbernya (source). Apache Spark adalah mesin pemrosesan data dalam memori, yang cepat dan ringkas dengan mode pengembangan API yang elegan dan ekspresif, yang memungkinkan pengguna

melakukan proses secara streaming, menggunakan pembelajaran mesin (machine learning), atau SQL yang memerlukan akses berulang-ulang secara cepat terhadap kumpulan data. Dengan Spark yang berjalan di Apache Hadoop YARN, developer sekarang dapat membuat aplikasi dengan memanfaatkan kehandalan dari Spark, untuk memperoleh wawasan, dan memperkaya data sains mereka dengan memasukkan data dalam satu kumpulan data besar di Hadoop.

4.3 Key Roles Proyek Analitik

Proyek Analitik Big Data melibatkan sejumlah peran kunci untuk memastikan keberhasilan dan efektivitasnya. Berikut adalah beberapa peran kunci yang umumnya terlibat dalam proyek Analitik Big Data:



Gambar 4.3 Key Roles Kunci Sukses Proyek Analitik

A. Business User

Business User: Seseorang yang memahami wilayah domain (kondisi existing) dan dapat mengambil manfaat besar dari hasil proyek analitik, dengan cara konsultasi dan menyarankan tim proyek pada scope proyek, hasil, dan operasional output (terkait dengan cara mengukur suatu variabel). Biasanya

yang memenuhi peran ini adalah analis bisnis, manajer lini, atau ahli dalam hal pokok yang mendalam.

B. Project Sponsor

Project Sponsor: Bertanggung jawab terkait asal proyek. Memberi dorongan, persyaratan proyek dan mendefinisikan masalah core bisnis. Umumnya menyediakan dana dan konsep pengukur tingkat nilai output akhir dari tim kerja. Menetapkan prioritas proyek dan menjelaskan output yang diinginkan.

C. Project Manager

Project Manager: Memastikan bahwa pencapaian utama projek dan tujuan terpenuhi tepat waktu dan sesuai dengan kualitas yang diharapkan.

D. Business Intelligence Analyst

Menyediakan keahlian dalam domain bisnis berdasarkan pemahaman yang mendalam mengenai data, indikator kinerja utama (KPI), metrik kunci, dan intelijen bisnis dari perspektif pelaporan. Analisis Business Intelligence umumnya membuat dashboard (panel kontrol) dan laporan dan memiliki pengetahuan tentang sumber data dan mekanismenya.

E. Database Administrator (DBA)

Set up dan mengkonfigurasi database untuk mendukung kebutuhan analitik. Tanggung jawab ini mungkin termasuk menyediakan akses ke database keys atau tabel dan memastikan tingkat keamanan yang sesuai berada di tempat yang berkaitan dengan penyimpanan data.

F. Data Engineer

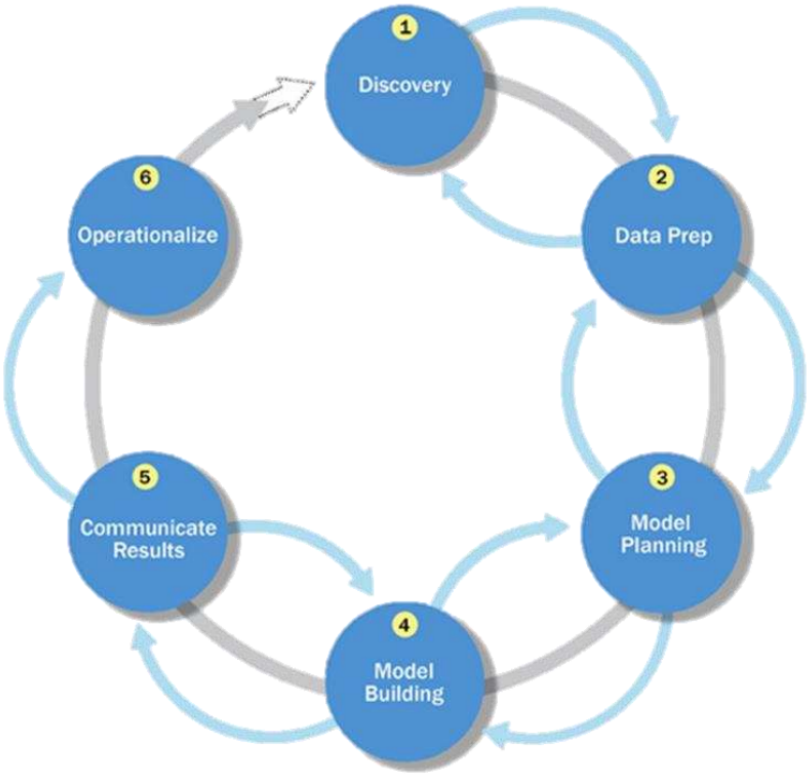
Memiliki keterampilan teknis yang mendalam untuk membantu penyetelan query SQL untuk pengelolaan data dan ekstraksi data, dan mendukung untuk konsumsi data ke dalam sandbox analitik. Data Engineer melakukan ekstraksi data aktual dan melakukan manipulasi data yang cukup besar untuk memfasilitasi kebutuhan proyek analitik. Insinyur data (Data Engineer) bekerja sama dengan ilmuwan data (Data Scientist) untuk membantu membentuk data yang sesuai dengan cara yang tepat untuk analisis

G. Data Scientist

Menyediakan keahlian untuk teknik analisis, pemodelan data, dan menerapkan teknik analisis yang valid untuk masalah yang diberikan. Memastikan melalui keseluruhan analitik tujuannya dapat terpenuhi. Merancang dan mengeksekusi metode analitis dan melakukan pendekatan lainnya dengan data yang tersedia untuk proyek tersebut.

4.4 Lifecycle Analitik Data

Lifecycle Analitik Data adalah serangkaian tahapan atau proses yang mencakup pengumpulan, penyimpanan, pengolahan, analisis, dan interpretasi data untuk mendapatkan wawasan yang bermanfaat. Lifecycle ini mencakup seluruh perjalanan data dari awal hingga akhir dalam konteks analitik. Meskipun setiap siklus hidup data bisa berbeda tergantung pada kebutuhan proyek atau industri, berikut adalah tahapan umum dalam Lifecycle Analitik Data:



Gambar 4.4 Gambaran Umum dari Lifecycle Analitik Data

Dari gambaran umum lifecycle analitik yang ditunjukkan dapat dilihat terdapat beberapa fase diantaranya sebagai berikut:

A. Fase 1 Discovery

Pada tahap ini, tim ilmuwan data (Data Scientist) harus belajar, mencari dan menyelidiki fakta-fakta, masalah (identifikasi problem base), mengembangkan konteks dan pemahaman, dan belajar tentang sumber data yang dibutuhkan dan yang telah tersedia untuk kesuksesan proyek analitik. Selain itu, tim merumuskan hipotesis awal yang nantinya dapat diuji dengan data. Tim belajar domain bisnis, termasuk kriteria dari data history yang relevan, seperti, apakah organisasi atau unit bisnis telah mencoba proyek serupa di masa lalu (apa saja

yang sudah mereka pelajari dari data). Tim menilai sumber daya yang tersedia untuk mendukung proyek tersebut dari segi SDM, teknologi, waktu, dan data.

Kegiatan penting dalam fase ini meliputi membingkai masalah bisnis sebagai tantangan analitik yang dapat dibahas dalam fase berikutnya dan merumuskan hipotesis awal (IHs) untuk menguji dan mulai mempelajari data.

B. Fase 2 Data Preparation

Tahap ini membutuhkan adanya sandbox analitik, di mana tim dapat bekerja dengan data dan melakukan analitik selama proyek tersebut. tim perlu melaksanakan proses ekstrak, load dan transformasi (ELT) atau ekstrak, transform dan load (ETL) untuk menyiapkan data ke sandbox. ETLT adalah proses integrasi data untuk mentransfer data mentah dari server sumber ke sebuah gudang data pada server target dan kemudian menyiapkan informasi untuk keperluan hasil akhir. Data Sandbox, dalam konteks Big Data, adalah platform terukur dan berkembang yang digunakan untuk mengeksplorasi informasi besar suatu perusahaan. Hal ini memungkinkan perusahaan untuk mewujudkan nilai investasi yang sebenarnya dalam Big Data. Sebuah sandbox data, utamanya dieksplorasi oleh tim Data Scientist yang menggunakan platform sandbox stand-alone, misal untuk analitik data marts, logical partitions pada suatu media penyimpanan di perusahaan. platform Data sandbox menyediakan komputasi yang diperlukan bagi para ilmuwan Data (Data Scientist) untuk mengatasi beban kerja analitik yang biasanya kompleks.

C. Fase 3 Model Planning

Dalam tahap ini tim menentukan metode, teknik, dan alur kerja. Mereka berniat untuk mengikuti tahap pembentukan model berikutnya. Tim mengeksplorasi data untuk belajar tentang hubungan antara variabel dan

kemudian memilih variabel kunci dan model yang paling cocok untuk digunakan.

D. Fase 4 Model Building

Tim mengembangkan dataset untuk pengujian (testing), pelatihan (training), dan tujuan produksi (menghasilkan data baru dari data yang ada). Selain itu, dalam fase ini tim membangun dan mengeksekusi model yang didasarkan pada kerja yang dilakukan di dalam fase Model Planning. Tim juga mempertimbangkan apakah ini alat yang ada akan cukup untuk menjalankan model, atau jika itu akan membutuhkan lingkungan yang lebih robust untuk mengeksekusi model dan alur kerja (misalnya, hardware yang cepat, teknik dekomposisi data dan pemrosesan paralel, jika dapat diterapkan).

E. Fase 5 Communicate Result

Tim bekerja sama dengan pemangku kepentingan (stakeholders) utama, menentukan apakah hasil proyek tersebut sukses atau mengalami kegagalan berdasarkan kriteria yang dikembangkan di Fase 1. Tim harus mengidentifikasi temuan kunci, mengukur nilai bisnis, dan mengembangkan narasi untuk meringkas dan menyampaikan temuan kepada para pemangku kepentingan.

F. Fase 6 Operationalize

Tim memberikan laporan akhir, pengarahan, kode, dan dokumen teknis. Selain itu, tim dapat menjalankan pilot project untuk menerapkan model dalam lingkungan produksi. Pilot Project adalah sebuah studi percontohan, proyek percontohan atau studi pendahuluan skala kecil yang dilakukan untuk mengevaluasi kelayakan, waktu, biaya, efek samping, dan efek ukuran dalam upaya untuk memprediksi ukuran sampel yang tepat dan memperbaiki desain penelitian sebelum kepada proyek penelitian skala penuh.

4.5 Kesimpulan

Arsitektur Big Data merupakan struktur konseptual yang mencakup lapisan-lapisan penting untuk menangani volume besar data. Komponen-komponen ini mencakup data source, data ingestion layer, data collector layer, data visualization layer, data query layer, data processing layer, dan data storage layer. Setiap lapisan memiliki peranannya masing-masing dalam siklus hidup data, memungkinkan pengelolaan data yang efisien dan efektif.

Proyek Analitik Big Data melibatkan sejumlah peran kunci yang berkontribusi pada berbagai tahapan proyek. Diantaranya adalah business user, project sponsor, project manager, business intelligence analyst, database administrator, data engineer, dan data scientist. Dengan peran-peran ini, proyek dapat dikelola dengan baik, dari pemahaman kebutuhan bisnis hingga implementasi solusi analitik yang efektif.

Siklus Hidup Analitik melibatkan serangkaian tahap yang memandu perjalanan data dari awal hingga akhir. Fasenya mencakup discovery, data preparation, model planning, model building, communicate results, dan operationalize. Dengan menjalani siklus ini, tim analitik dapat memahami data, membangun model yang relevan, berkomunikasi hasil secara efektif, dan mengintegrasikan solusi analitik ke dalam operasional bisnis.

Keseluruhan, pemahaman arsitektur Big Data, pengenalan peran-peran kunci dalam proyek analitik, dan penerapan siklus hidup analitik menjadi landasan penting dalam memahami dan mengelola data secara efektif dalam konteks Big Data. Dengan sinergi antara arsitektur yang baik, tim yang terampil, dan siklus hidup yang terorganisir, organisasi dapat mengambil manfaat maksimal dari potensi analitik Big Data.