Nama : Helmi Saifullah Mansur

Kelas : UGM07

Universitas Host : Universitas Gadjah Mada

Tugas Mandiri Sesi Industri

**Sharing Session on Data and Analytics**

* Dalam pengembangan AI dibutuhkan landasan Data Science dimana keduanya juga merupakan bidang yang saling terkait. Hal ini perlu disiapkan untuk menghadapi era AI kedepannya
* Untuk data scientist nantinya akan mengembangkan model terbaik dari data untuk dapat menjawab permasalahan bisnis yang dihadapi.
* Dalam data science ada beberapa proses yang harus dijalani yaitu sebagai berikut:

1. Business Understanding : tahapan menentukan masalah, tujuan, dan solusi dari perspektif bisnis
2. Analytic Approach : pendekatan dari masalah yang ditemukan dengan teknik descriptive, predictive, atau perspective
3. Data Requirement : mendefinisikan data - data yang dibuthkan dalam pemrosesan
4. Data Collection : proses pengumpulan data sesuai masalah yang ada
5. Data Understanding : proses visualisasi data yang dilakukan untuk membantu memahami kualitas data
6. Data Preparation : proses eksplorasi data bisa dengan melakukan cleaning, Transforming, selection, dan normalize
7. Modeling : proses memilih dan membuat model yang optimal untuk data yang diproses dengan algoritma machine learning. Seperti clustering, regresi, klasifikasi atau rekomendasi
8. Evaluation : proses penilaian tingkat keoptimalan proses modelling bisa dengan matrix atau confussion matrix
9. Deployment : Proses implementasi dari serangkaian kegiatan atau proses yang sudah dilakukan diatas

* Data Lake vs Data Warehouse

|  |  |
| --- | --- |
| Data Lake vs Data Warehouse | |
| Data Lake | Data Warehouse |
| Data unstructured | Data structured |
| Users : Data Scientists, Data Analysts | Users : Business Analysts |
| Use cases : Stream Processing, Machine Learning, Real Time Analysis | Use cases : Batch processing, BI, Reporting |
| Data Lake contain minimal pre-processing and can be used contain unconventional data | Data warehouse is cleaned, pre-processed and refined |
| Data Lake can be in any forms or size, large amounts of unstructured data can be stored indefinitely and can be transformed when is use only | Data Warehouse contain less data in order of terabytes |
| Data in data lake can be used for a wide variety of applications | Data Warehouse contain historic and relational data |

* Kind of skillsets are required in data engineering :

1. Become proficient at programming
2. Learn automation and scripting
3. Understand our databases
4. Master data processing techniques
5. Schedule our workflows
6. Study cloud and computing
7. Internalize infrastructure
8. Follow the trends

* Various learning sources available online :
* Free Resources : Kaggle, Coursera, Youtube
* Paid Resources : DataCamp, Udemy,Udacity
* We can learn dashboarding using drag and drop with data visualization tools like Power BI, Google Data Studio and Tableau