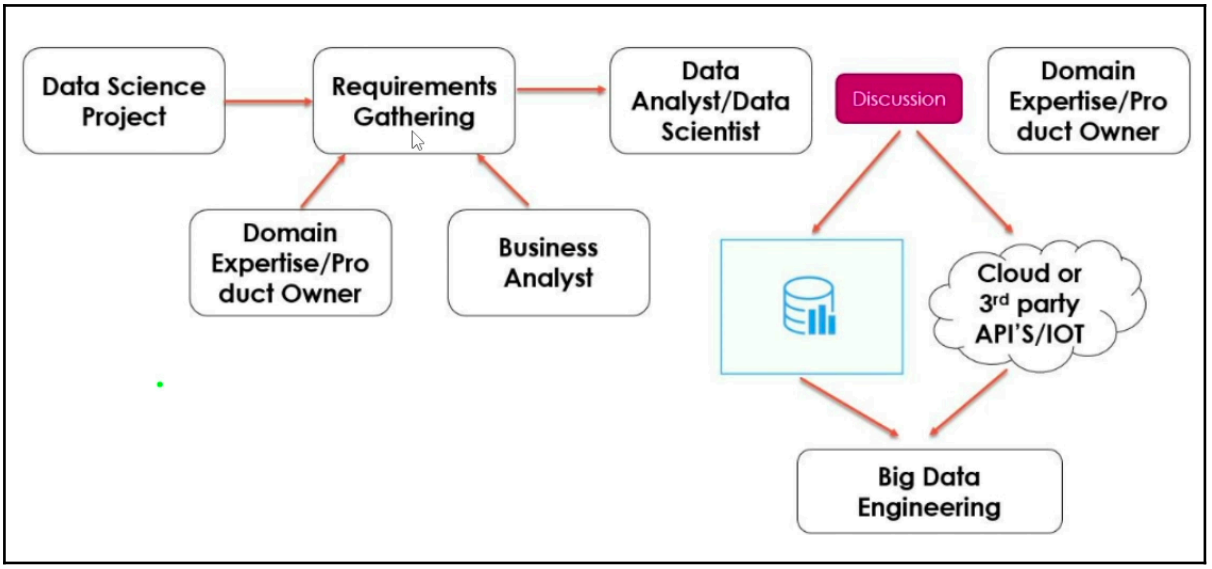
Data Science Life Cycle (DSLC)



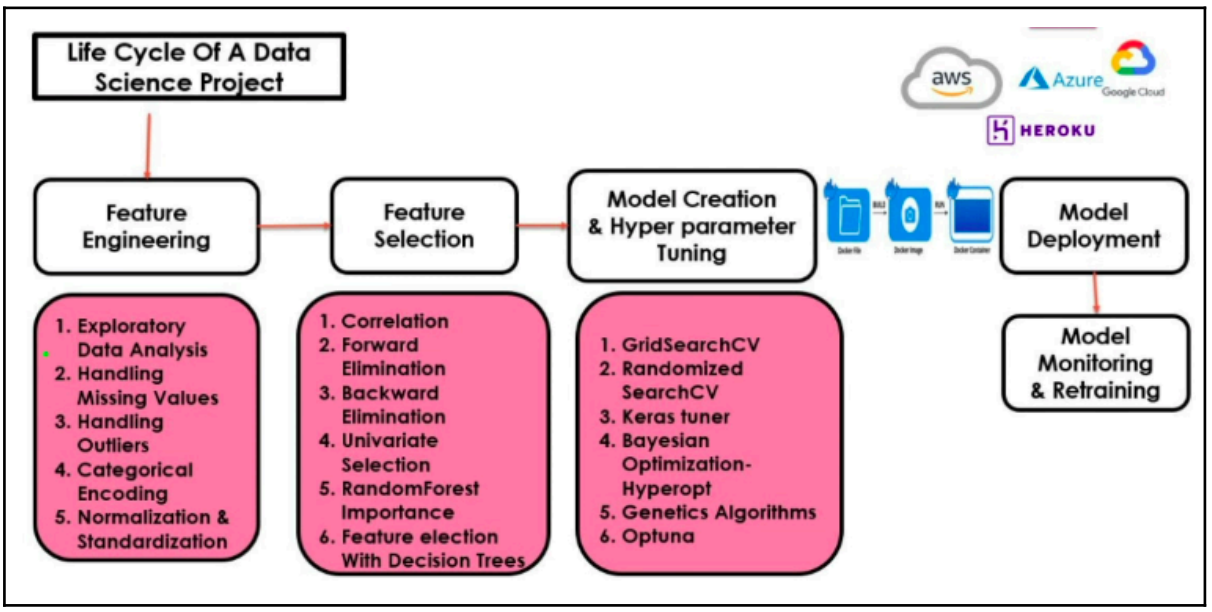
In a data science project, the first step is requirement gathering, for this we have domain experts or product owners along with business analysts. They will know what data science project they really want to implement. They will write down all the requirements, what all stories need to be completed and they will try to divide this entire story in the form of sprints as in data science projects most companies follow an agile process. Each and every one of these sprints will try to divide the requirements like which all modules need to be developed in sprint 1 followed by sprint 2. Once all the requirements are basically written down, then all of them will be sent to the data analyst or data scientist team.

Again this data analyst and data science team will have a discussion with the domain expertise or product owner to find out what all data will be required in order to solve this particular project. Basically together they will try to identify the data or the source of the data. The data may be present in the internal database, present in the third party or cloud API and it may also be an IOT data.

Once the data or the source of the data is identified then the big data engineering team will come into picture which will be responsible for creating a data pipeline. The data pipeline will be created in such a way that all of the identified data sources will be combined in the form of a data pipeline. One of the most common data pipelines in the industries that are done by the big data engineering team is the ETL pipeline which is

nothing but Extract, Transform and Load from multiple data sources. Transformation is done and finally loads the entire combined data into some data source like MongoDB, PostgreSQL etc.

Overall, big data engineers create amazing data pipelines wherein they integrate with multiple data sources and then try to combine the data and load it into a specific data source (MongoDB). Once the entire data is loaded in the MongoDB, this will keep on getting updated on a daily basis or on a weekly basis depending on the data that is probably coming from the third party APIs or any other data source. Once this is done then our life cycle of a data science project comes.



Whatever data has been collected in the data source (MongoDB) which will be hosted in the cloud. Now from MongoDB, our life cycle of a data science project.

A step before Feature engineering (First step in data science life cycle) is **data ingestion** followed by feature engineering, feature selection and all. The below are the steps in the data science Life cycle :

1. Data Ingestion
2. Feature Engineering
3. Feature Selection
4. Model Creation and Hyper parameter tuning
5. Dockerization
6. Model deployment
7. Model Monitoring & Retraining