Data Engineering

Data Engineering Concepts

CTO of Nexart decided to merge Data Science capabilities, Machine Learning and AI with their generated data to increase revenue and to anticipate future revenue losses of the organization. AI’s are empowered with quality data. Machine Learning algorithms analyze the data and make predictions. The principles of Data Engineering guide in the Data Analysis processes. Data Science and Data Engineering fields have common characteristics. The picture given below demonstrates aspects in IT where data science is applied.

The data engineers of Nexart gather generated data from departments and store in the MongoDB database for data scientists, data analysts, business intelligence developers and other IT specialist within the Nexart to access at any time. They choose the MongoDB database due to it’s scalability. The image given below demonstrates the steps that are followed in data engineering process.

All the collected data from Nexart Departments will be moved to MongoDB database in the data ingestion state. In the data transformation stage, highly demandable data will be cleaned, normalized and converted to required format. The users of Nexart that requested data will be served with data in the data serving process. The transparency of data engineering process is maintained through data flow orchestration. This gives the privilege of monitoring data workflows and repair data quality and performance issues. Automating the steps followed in the data engineering process is defined as a data engineering pipline. Data from the MSSQL database is moved to MongoDB database with aid of data engineering piplines by coupling tools with operations. In addition data engineers of Nexart could write a script to get weekly/monthly/yearly sales reports. Hence the generation of sales reports is a repetitive task.

Since Nexart enterprise is dealing with big data, the data engineers decided to go with a Big data engineering pipeline that uses ELT method. The image below captures out the whole ELT process.

ETL functions differently compared to ELT pipelines. We can observe the functionality of ETL from the image given below.

Apart from data migrations, Nexart can deploy the pipeline for data wrangling, data integration and copying databases.A data pipline consists of 7 components and they are origin, destination, dataflow, storage, processing, workflow and monitoring. MSSQL and MongoDB paly the role of source and destination respectively. The dataflow defines changes that data undergo from the source to the destination. As data move from MSSQL to MongoDB, the data will be stored in the lake.

Processing defines the ingestion, storing and transformation of data from the source to the destination. The workflow defines of set of activities that are carried out from the Source to the destination and it also highlights dependencies that each task have on them. The monitoring mechanism ensures the data is in proper shape. It would be ideal for Nexart to maintain two data pipelines. One pipeline should follow batch processing while the other should stream processing. In some instances, Nexart would have to look into old records make a business decision and sometimes the corporation would have to deal with real-time data such as data generated from robots at production line. So in this case it’s better to maintain two data pipelines.

Data pipeline tools branches under ELT, Data Warehouing, Data Lake, Batch Work-flow Schedulers, Real-Time Data Streaming and Big Data tools. All the data preparation and integration tools falls under ETL tools and IBM DataStage Informatica power Center, Oracle Data Integrator, Talend Open Studio are examples for ETL tools. Instances for Data Warehouse tools are Amazon Redshift, Azure Synapse, Google BigQuery, Snowflake and Teradata. AWS, Microsoft Azure, Google Cloud and IBM are examples for cloud service providers which offer data lakes as tools. Luigi and Azkaban are batch workflow schedulers which declares tasks programatically along with their dependencies. In addition they have the capability of monitoring and automating the tasks. Apache Kafka, Apache Storm, Google Data Flow, Amazon Kinesis, Azure Stream Analytics, IBM Streaming Analytics and SQL Stream are tools that are used process real-time data that are generated by machine sensors, IoT sensors etc. The data pipilne tools mentioned do not go well with Big Data. The software developers have developed specific tools to perform heavy duty operations such as prepping data pipelines for big data. Let’s review each of these tools. Hadoop and Spark platforms are used batch processing big data. The Spark streaming analytics service tool enhances the functionality of the Spark platform. Apache Oozie and Apache Airflow tools are used by data engineers to perform scheduling and monitoring for batch jobs. Cloud service providers such as Amazon, Google, IBM, Microsoft and Alibaba offer their own tools for data experts to create big data pipelines. As we have acknowledged the DNA of data engineering, the next segment discusses how of the principles of data engineering is applied towards the Nexart case study.

Data Engineering Pipelines

As we acknowledge before Nexart decided to go with the ELT method for both data pipelines. The data engineers of Nexart is not sure about the way of transforming data and they have an idea of generating insights in the future as well. Due to these reasons, they decided to go with the ELT method. Since Nexart is generating tons of data daily from the production line and other departments, they decided to go with tools that are specifically designed for big data. Let’s discuss about the data pipelines further.

In order to increase the revenue, Nexart needs valuable insights from data that been stored in the MSSQL database. These insights are been generated using machine learning. Models need to be fed loads of data to get better insights. So the model requires processed data from the past and the present. So they decided go with a batch processing data pipeline to get business insights. Apache NiFi, MongoDB, Apache Spark and Airflow tools will be used for the batch processing data pipeline. In this pipeline there are three phases. Data stored in the MSSQL database needed to extracted in the first hand. Before commencing the first phase, the system engineers of Nexart will determine the peak hours of service by analyzing incoming and outgoing traffic of Nexart. After that the data engineers of Nexart will extract data from MSSQL database according to off peak hours defined by the data engineer using Apache NiFi software. This application gives the privilege of setting up data extraction schedules. As a plus point this doesn’t clash the ongoing services of Nexart during peak hours. Extracted data will be needed to store in a storage. The schemaless characteristic of MongoDB database addresses the storage problem by storing all extracted raw data from MSSQL database. This enables to maintain the integrity of data. These raw data doesn’t have a meaning. So these data needed transformed to get a better meaning. As a result transformation takes place using Apache Spark software. All the transformations jobs will be carried using Spark. Repetitive transformations such as generating weekly reports can be hectic sometimes. As a remedy all the transformation will be automated using Airflow application. Batch processing pipelines are very useful in generating weekly/month/yearly sales reports, salary reports etc.

According to the golden rule of information, the value of data that is being generated stays constant for a certain period of time and the value of information starts to fade away with time. Robots in the production generate piles of data daily using sensors in real-time. These data may have a significant value. So it would be unwise to ignore the data. Sine the data experts are dealing with real-time data, they decided to deploy the stream processing method for the second pipeline. The ELT technique is used for the second pipeline due to the benefits that are offered. This pipeline consists of three phases as well. The data that is been produced by the sensors of robots will be captured by the Apache Kafka software as events. Next the captured data will be stored in the MongoDB database in the loading phase. In order to transform the raw data to meaningful data, Apache Spark application will be used. In some case scenarios robots used to deviate from it’s normal operational task. This results in downtimes or product defects. A stream pipeline can solve this issue. As Apache Kafka extracts and stores data MongoDB database, Apache Spark will the raw data and detect for any anomalies in the real-time data. If the application detects an anomaly it will alert the relevant IT professionals to about the event. So that the IT professional could take the necessary steps troubleshoot any issues in the robot or production line. The data migration timeline is divided into three stages and they are pilot phase, full rollout and post migration optimization. In the pilot phase a small of data is migrated into MongoDB and tested. If the tested data are valid, all the data are migrated into the database in full rollout stage. Additional tinkering to the data are performed in the post migration optimization phase. Since Nexart have choose MongoDB to store data, let’s dig into the roots of MongoDB from the next section.