24/02/2024

CI/CD

① What is CI/CD?

→ A CI/CD pipeline is a concept central to software. It spans a whole field of processes, testing methods, and tooling, all facilitated by the Git code versioning process.

→ CI checks and tests every new piece of code (or data transformation logic) you add to your data pipeline.

→ CD ensures that once tested and approved, this code gets added to the live system without manual intervention.

→ CI, CD, in the context of data pipeline deployment, focuses on automating data operations and transformations. This merges development, testing and operational workflows into a unified, automated process, ensuring the data assests are consistently high quality and the data infrastructure evolves smoothly, even at scale.

→ Using CI/CD for data pipelines automation has become more critical in ensuring the development velocity of processes such as training machine learning models, supporting a data science team, doing large-scale data analysis, business intelligence or data visualization, supporting the growth of unstructed data collection, and other business needs.

Continuous Integration (CI) in data pipelines.

① Automated testing:-

Automated tests check the integrity and quality of data transformations, ensuring the data is processed as expected and any error is spotted early.

② Version Control:-

Data pipeline code (eg:- SQL scripts, Python transformations) is stored in repositories like Git, allowing tracking and managing

changes.

③ Consistent Environment:-

CI tools can run test in environments that mirror production, ensuring that differences in configuration or dependies don't introduces errors

④ Data Quality checks:-

These might includes for null values, data range violations, data type mismatches or other custom quality rules.

Continuous data pipelines deployments

① Automated Deployment:-

Once code change pass all CI checks, CD tools can automate their deployment to production, ensuring seamless data flow.

② Monitering and Alerts:-

Once deployed, monitoring tools keep track of the data pipelines performance, data quality, and any potential issues. Automated alerts can notify on discrepancies.

③ Roll backs:-

In case an issue is identified post-deployment, CD processes allow for quick rollbacks to a previously stable state of the data pipeline.

④ Infrastructure as Code (IaC):-

Many CD tools support IaC's. For example, cloud resources such as storage or compute can be provisined automatically as part of the deployment process.

Git :-

→ Git is a distributed version control system that facilitates collaborative software development by tracking changes across multiple contributions.

→ Git can be paired with data orchestration tools and integrated into CI/CD workflows, providing the benefits of streamlined deployment and consistency in data engineering tasks.

ETL pipelines :-

→ ETL (Extract, Transform, Load) pipelines are at the heart of the data engineering. They're the processess that pull data from source (databases, API's, etc..), transform it into a usable format, and then load it into a destination, like databases or a data warehouse.

when you deploy an ETL script to Git, you're not just saving the code - you could be triggering a series of events:

① Testing :-

Automated tests are first run to ensure the new code doesn't break anything.

② Deployment :-

Once tests pass, the ETL processes can be automatically deployed to a staging or production environment.

③ Notifications :-

If any part of the process fails, or if it's successfully completed, notifications can be sent out.