Assessment 1

CID 01252821

Data Pipelines

Introduction to Data Pipelines

- What is a data pipeline?: definition, a brief overview, including a diagram
- Main purpose: encompasses the ways data flows from one system (source) to another (destination)
- Energy sector context:
  - Improving decision-making and operational efficiency within the energy sector
  - How: data pipelines can handle diverse data types from various sources, such as sensors on energy grids, customer usage records, and environmental impact data, leading to efficient data processing and analysis, and hence fastening informed decisions made by the energy company.
  - Possible use cases: support of predictive maintenance, demand forecasting, and the integration of renewable energy sources into the grid

Main Components of Data Pipelines

Overview of the main components of data pipelines (following the definition and diagram form above):

- Data ingestion (source): collection of data from various sources, types (batched, streaming)
- Data processing: transformation, cleaning, normalization (of sensor data), frameworks for doing so depending on the nature of the data
- Data storage (destination): different types of storage depending on whether the data is structured or unstructured (RDBMS, NoSQL, data warehouse, data lakes), importance of selecting the right type of storage in terms of scalability, performance, and complexity and specific use cases needs (data structure, mutable vs immutable, faster reading vs faster writing), might be the same as source
- Data visualization and analysis: final component involving analyzing the stored data and presenting it in a user-friendly format through visualization tools (transforms data into insights helping the decision-making processes), types of visualizations, tools, integrations
- Mention specific pipeline structures such as ETL, ELT

Challenges and Solutions

- Data quality:
  - Typical problems: missing data, outliers, data values that violate specific threshold values (positive values, negative values, specifc use case thresholds), wrong data types
  - Why are these problematic?: may lead to inaccurate analytics, misleading insights, and potentially costly business decisions
  - Solution: automated runtime data quality checks, including on input and on output so bad data cannot reach the destination, possible frameworks such as great_expectations
- Handling large volumes of data: right choice of data storage, frameworks
- Orchestration
- Maintaining pipeline performance: crucial for timely data processing and analysis
- Broken data pipelines:
  - Reasons: failure within the data pipeline, failure independent of the pipeline - platform errors, database errors, etc.

  - Monitoring and alerts: can prevent significant data processing delays or losses

Extend data pipelines to machine learning pipelines

- Integration of AI and machine learning models into data pipelines for advanced analytics (Machine Learning pipelines, MLOps)
- Integrate predictive maintenance, demand forecasting in the data pipelines

Why are data pipelines ethical?

- Ensuring fairness, transparency, and accountability
- Transparency (a core ethical principle): openly documenting the data sources, methodologies, and algorithms used across data pipelines, leading to help in building trust with customers, regulators, and stakeholders by demonstrating a commitment to fairness and ethical use of data, especially in decisions affecting pricing, service availability, or investment in sustainable resources
- Reproducibility: essential for ensuring that the models and analyses the company relies on can be recreated and verified for accuracy over time, and that the company's operations adapt to changing market conditions and technological advancements.
- Accountability: requires the company to maintain a comprehensive audit trail for data pipelines and decisions made by AI or ML models (helpful when addressing any disputes from customers regarding billing, services)

Resources

- Densmore, J. (2021). Data pipelines pocket reference : moving and processing data for analytics. O'Reilly
- Kleppmann, M. (2017). Designing Data-Intensive Applications. O'Reilly Media.
- https://www.ibm.com/topics/data-pipeline
- https://www.databricks.com/glossary/data-pipelines
- https://aws.amazon.com/what-is/data-pipeline/
- great_expectations: https://github.com/great-expectations/great_expectations
- Ameisen, E. (2020). Building Machine Learning Powered Applications: Going from Idea to Product. O'Reilly Media
- https://www.fivetran.com/learn/data-pipeline-vs-etl

Homomorphic Encryption

Introduction to Homomorphic Encryption

- What is homomorphic encryption? - definition, brief overview, types of homomorphic encryption
- Main purpose: allows computations to be performed on encrypted data without first having to decrypt it
- Energy sector context:
  - analysis of users' encrypted smart meter data in the cloud to optimize grid performance without exposing users' identities
  - securely forecasting energy demand from encrypted consumption patterns without compromising customer privacy
  - enhancing predictive maintenance on encrypted operational data from energy assets, ensuring confidentiality

Types of Homomorphic Encryption

Types of homomorphic encryption: definition, use cases, pros and cons

- Partial Homomorphic Encryption (PHE)
- Somewhat Homomorphic Encryption (SHE)
- Fully Homomorphic Encryption (FHE)

Homomorphic Encryption in Machine Learning

Use Cases:

- Secure Data Sharing
- Privacy-preserving Predictive Analytics
- Federated Learning

Challenges:

While the integration of homomorphic encryption into machine learning offers considerable benefits for data privacy, it also introduces several challenges:

- Computational Complexity Limitations
- Model Complexity Limitations

Solutions:

- Algorithm Optimization
- Hardware Acceleration
- Simplified Models and Techniques

Note: Might focus specifically on Fully Homomorphic Encryption.

Why is Homomorphic Encryption ethical?

- Enables the secure processing of sensitive data while preserving privacy and confidentiality
- Allows for data to be encrypted and remain so during analysis, ensuring personal information is protected against unauthorized access and breaches
- upholds the ethical principles of respect for privacy, data protection, and trustworthiness in data handling practices

Resources

- Gentry, C. (2009). A Fully Homomorphic Encryption Scheme. Stanford University.
- Gentry, C. Computing Arbitrary Functions of Encrypted Data. IBM T.J. Watson Research Center

- https://en.wikipedia.org/wiki/Homomorphic_encryption
- Chen et al. (2018). Logistic regression over encrypted data from fully homomorphic encryption
- Graepel et al. (2012). ML Confidential: Machine Learning on Encrypted Data.
- Iezzi (2020). Practical Privacy-Preserving Data Science With Homomorphic Encryption: An Overview.

Additional materials

Data pipeline diagram proposal:

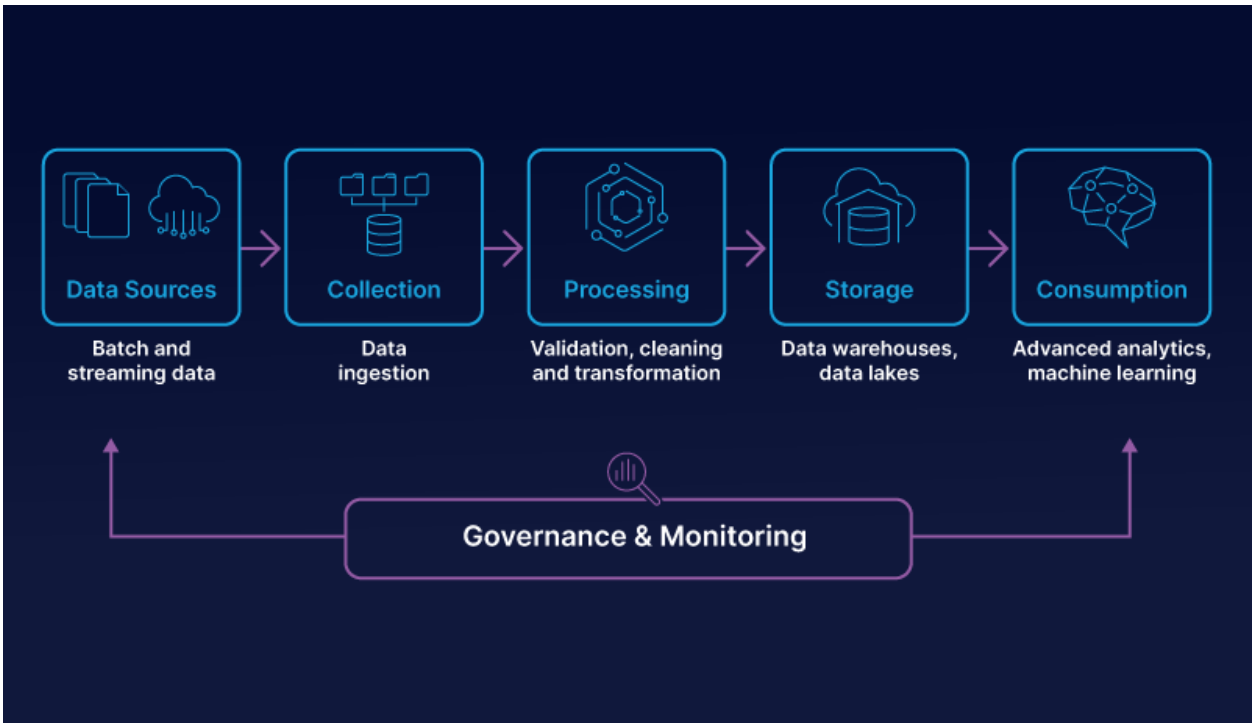

Figure 1: Source: https://www.striim.com/blog/guide-to-data-pipelines/