MSc Data Science

DAT 608:
Big data technologies

Dr. Pius Onobhayedo

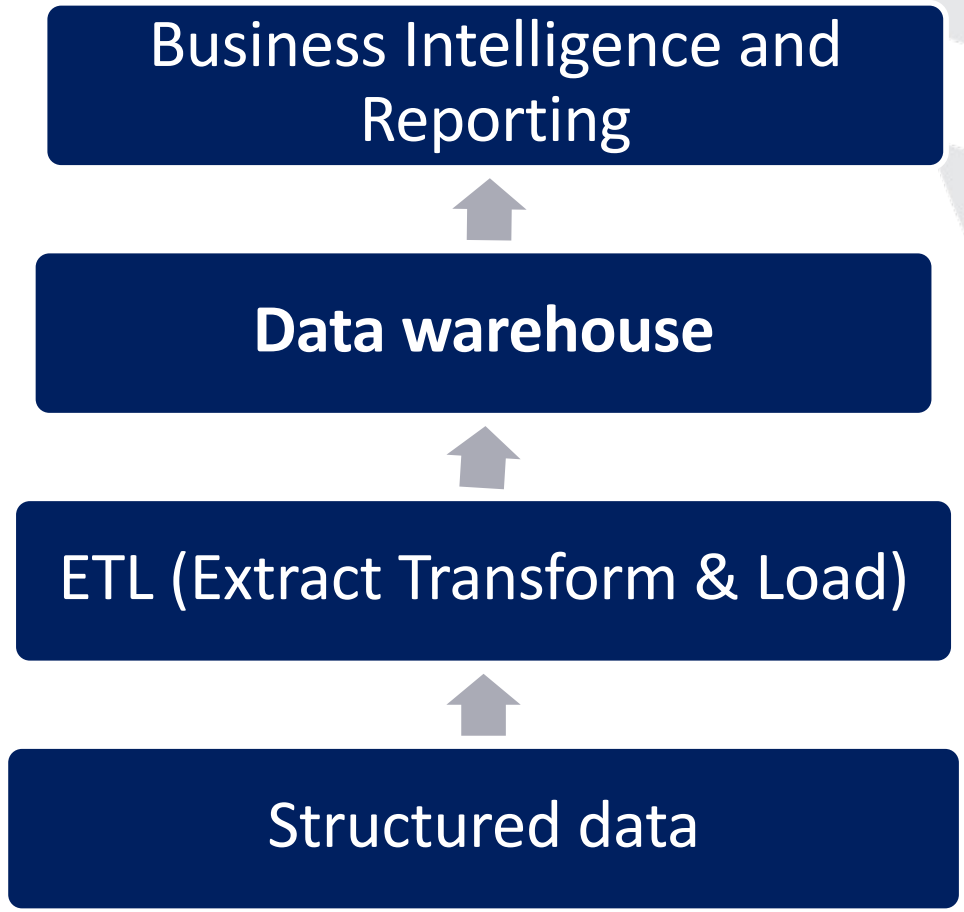
ponobhayedo@pau.edu.ng

Outline

- Section A: Introduction
  - Overview of Big data concept
  - Sources of data
  - Challenges with working on big data

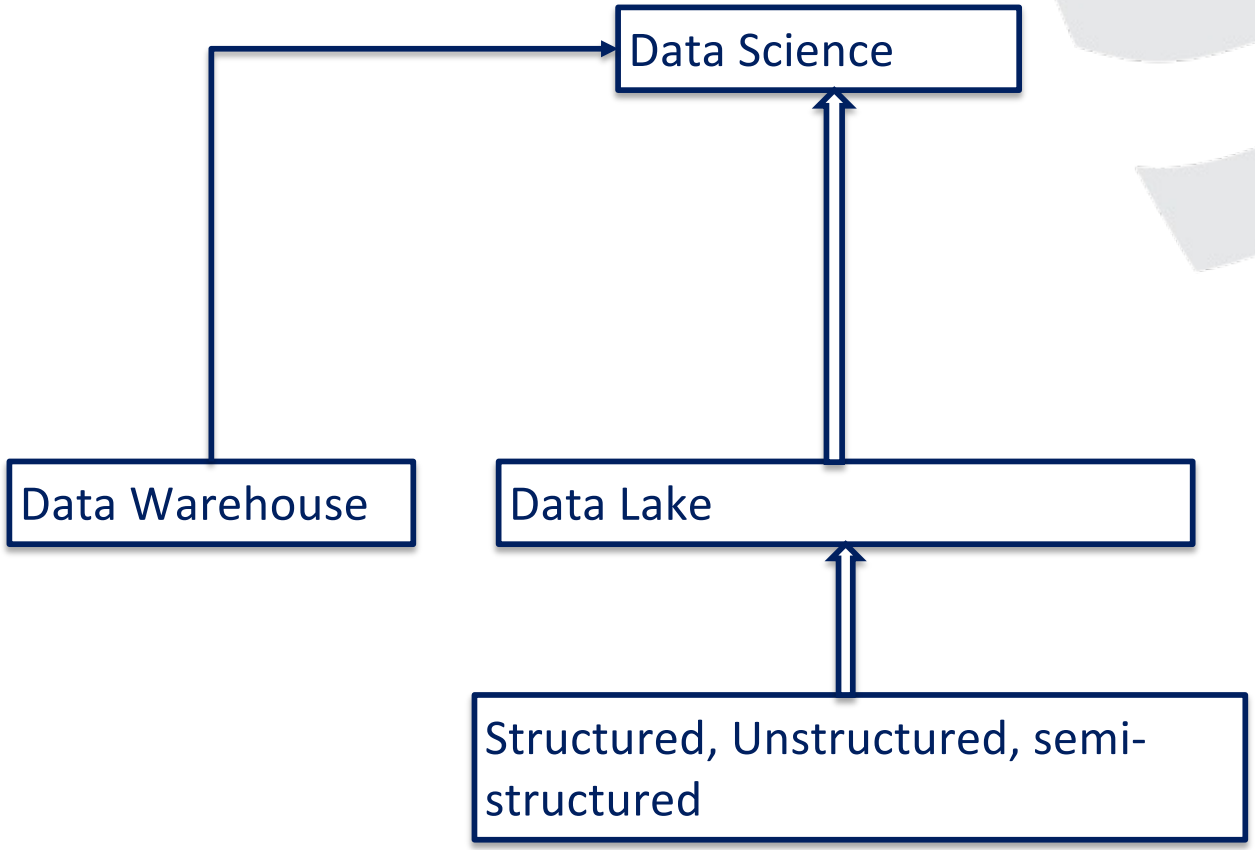
…Outline

Section B: Data Warehousing



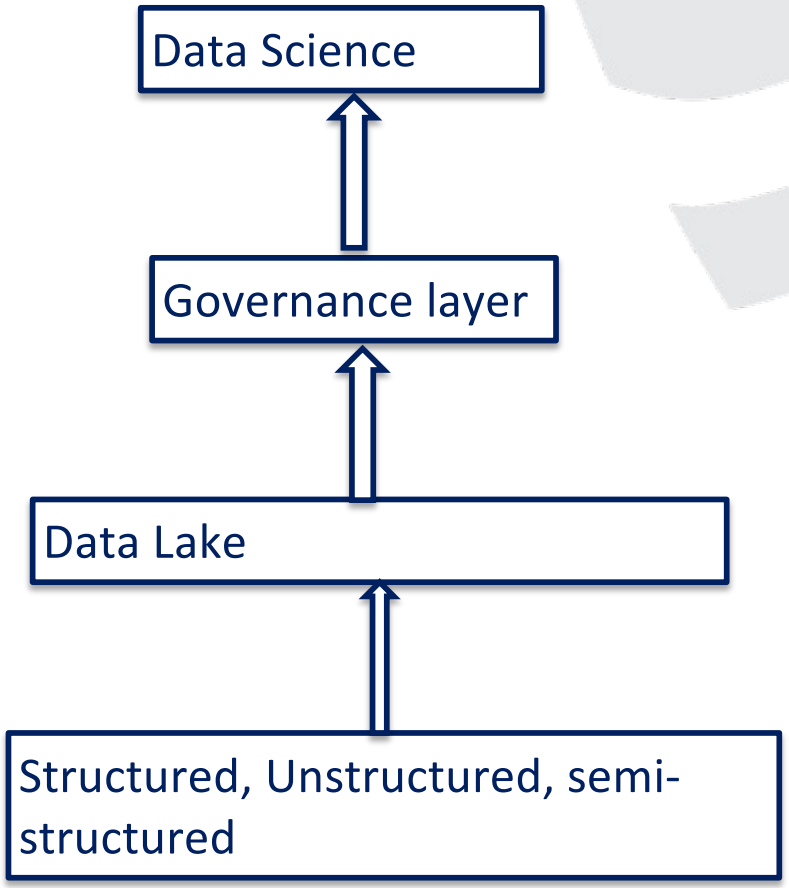
…Outline

Section C: Data Lake



…Outline

Section D: Data Lakehouse



…Outline

- Tools to be explored:
  - CockroachDB for resilient distributed data storage
    - https://www.cockroachlabs.com/
  - Streaming data ingestion and ETL with Apache Kafka
    - https://kafka.apache.org/
  - ETL with Kafka
  - Apache Spark for big data processing
    - https://spark.apache.org/

…Outline

- …Tools to be explored:
  - Delta Lake for Lakehouse
    - https://delta.io/
  - Trino for scalable queries
    - https://trino.io/
  - Cube for organized caching and API endpoints
    - https://cube.dev/blog/cube-integration-with-trino-sql-query-engine-for-big-data

…Outline

- …Tools to be explored:
  - ipfs for blockchain related data
    - https://ipfs.tech/