Introduction to Data Engineering

Janaka Wijayanayake

Course Aim/Intended Learning Outcomes:

At the completion of this course student will be able to:

- Differentiate data engineering from data science.
- Reason why data engineering as an important emerging discipline.
- Explain the critical activities in data engineering
- Use the commonly used software tools to create data pipelines and automate the underlying processes.

Evaluation Criteria

Quiz	30%
Topic Presentation	20%
Research Paper Presentation	20%
Project Presentation	15%
Project Report	10%
Project Proposal	05%

Lecture Outline

- Data, Big Data and Challenges
- Data Engineering
  - Introduction
  - Why Data Engineering
- Data Engineer
  - What do they do?
- Major/Concentration in Data Engineering

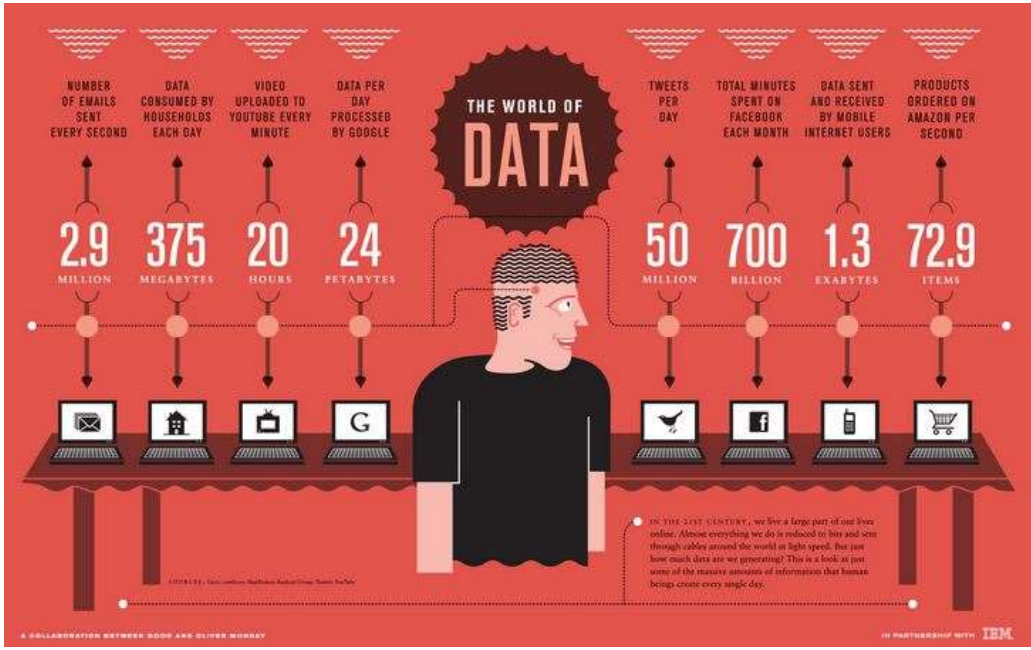
Data All Around

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network

How Much Data Do We have?

- Google processes 20 PB a day
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day
- 1000 genomes project: 200 TB

- Cost of 1 TB of disk: $35
- Time to read 1 TB disk: 3 hrs
  (100 MB/s)

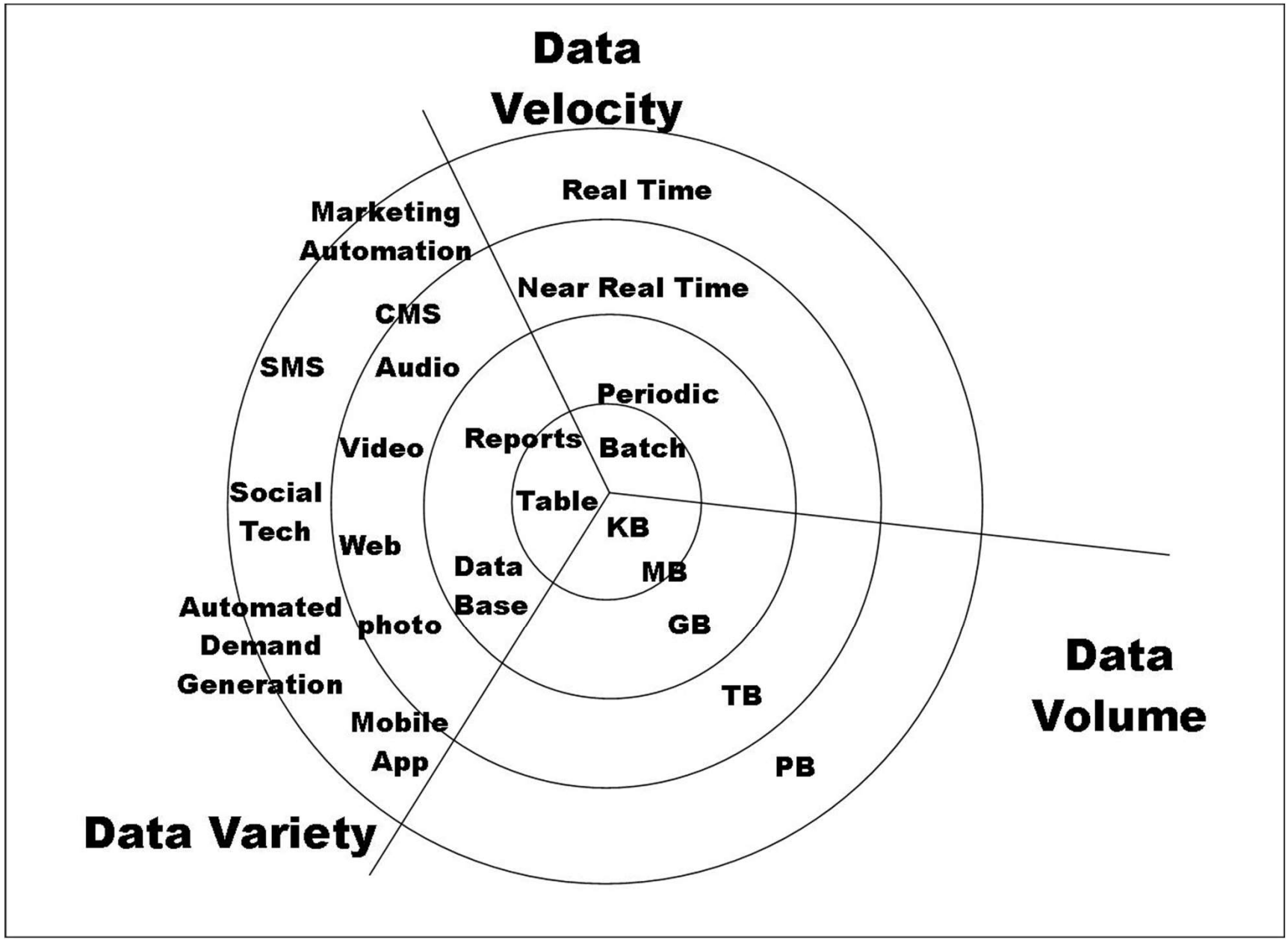


Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

  - 5 Vs - Volume, Velocity, Variety, Value, and Veracity
  - 7 Vs - Volume, Variety, Velocity, Value, Veracity, Variability and Visualization

Big Data


Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Table
Social
Tech
Web
KB
MB
Data
Base
GB
Automated
Demand
Generation
photo
TB
Mobile
App
PB
Data
Volume
Data Variety

Types of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data

What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

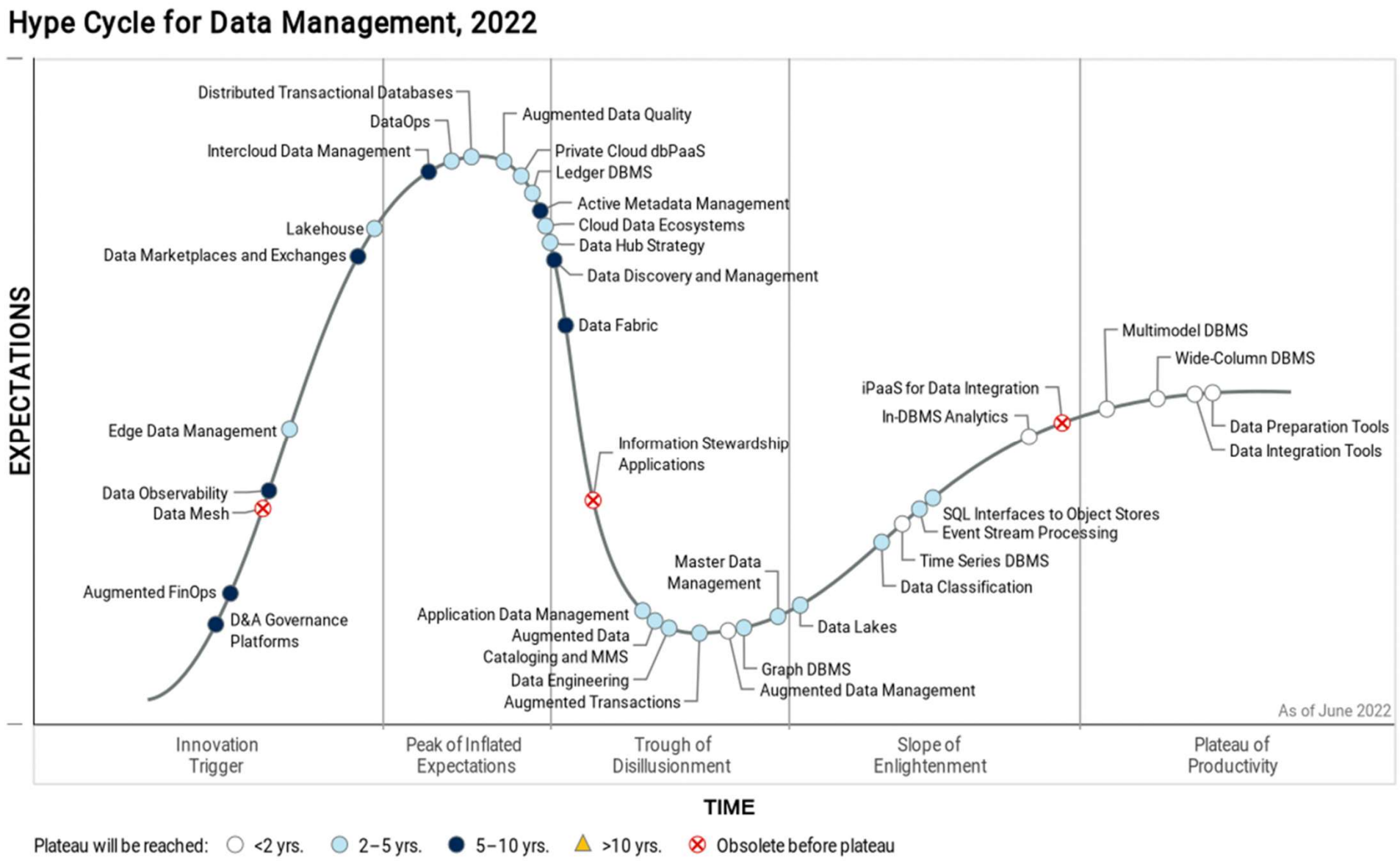
What is Data Engineering?

- An area that manages, manipulates, extracts, data from large amount of data
- Data Engineering (DE) is a multidisciplinary field of study with goal to address the challenges in big data
- Data Engineering principles apply to all data – big and small

What is Data Science?

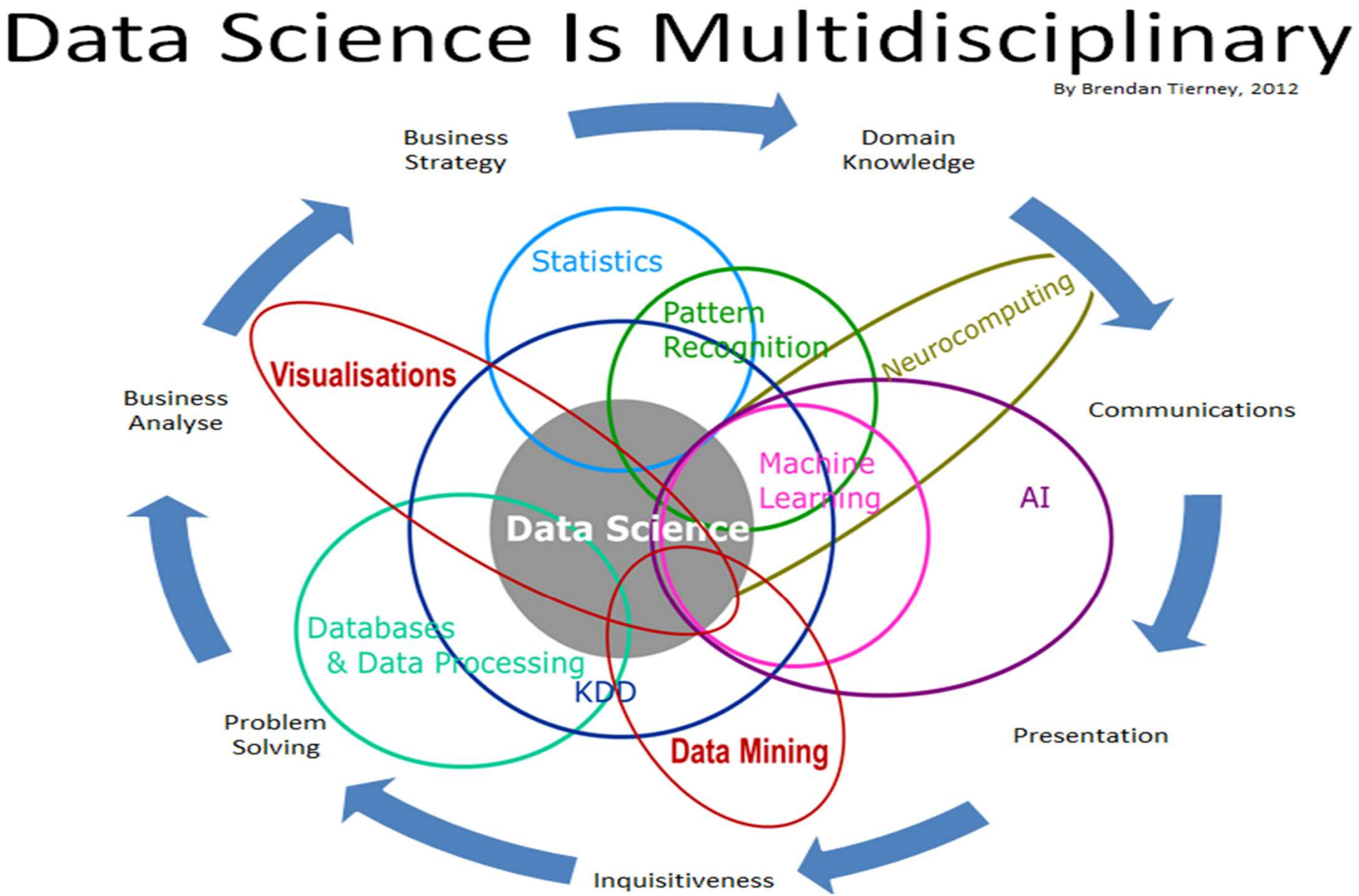
- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.
  - Operations Research
    - Optimization

Figure 1: Hype Cycle for Data Management, 2022

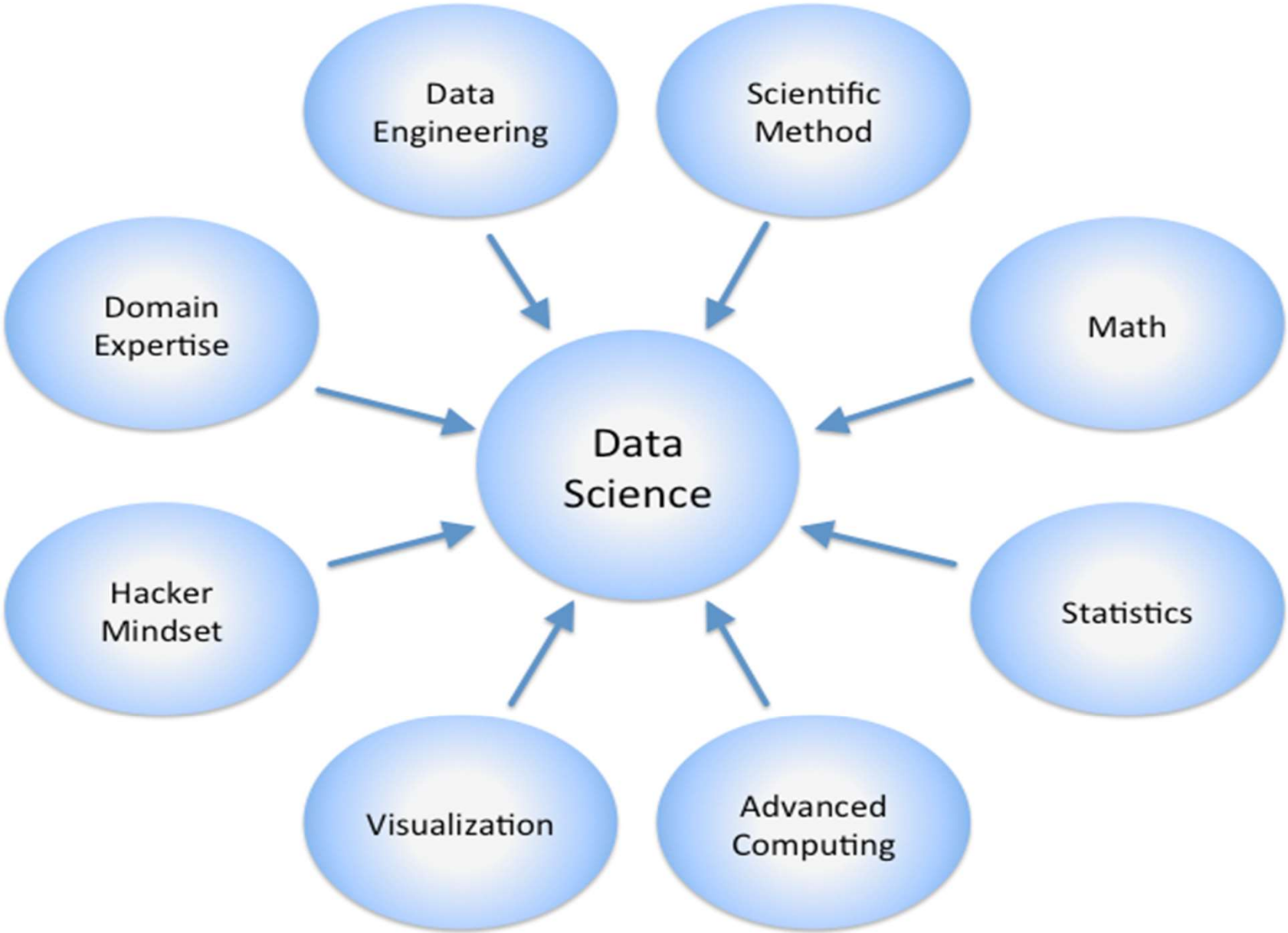


Source: Gartner (June 2022)

Data Science



Data Science



Data Scientists

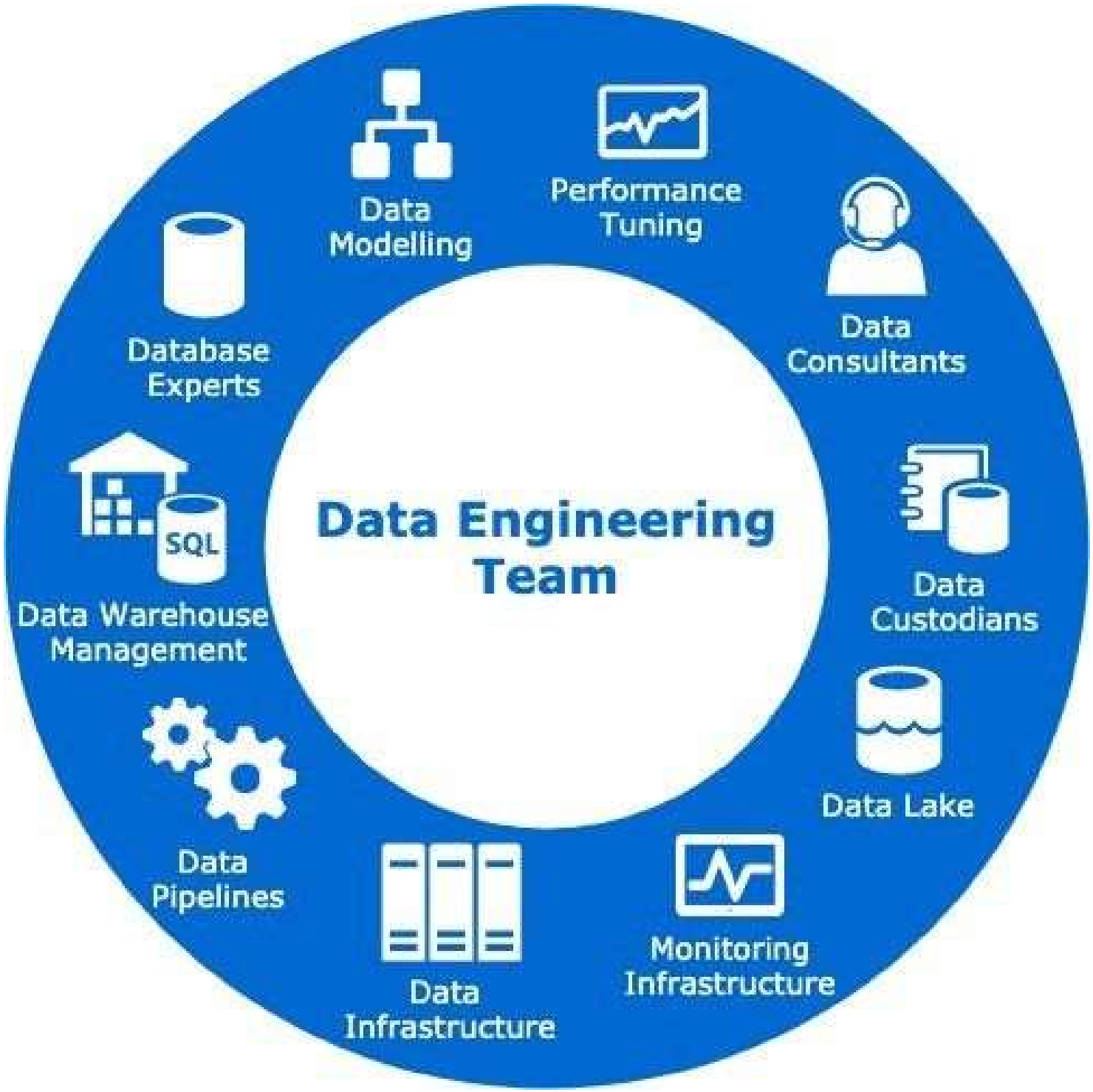
- They find stories, extract knowledge. They are not reporters

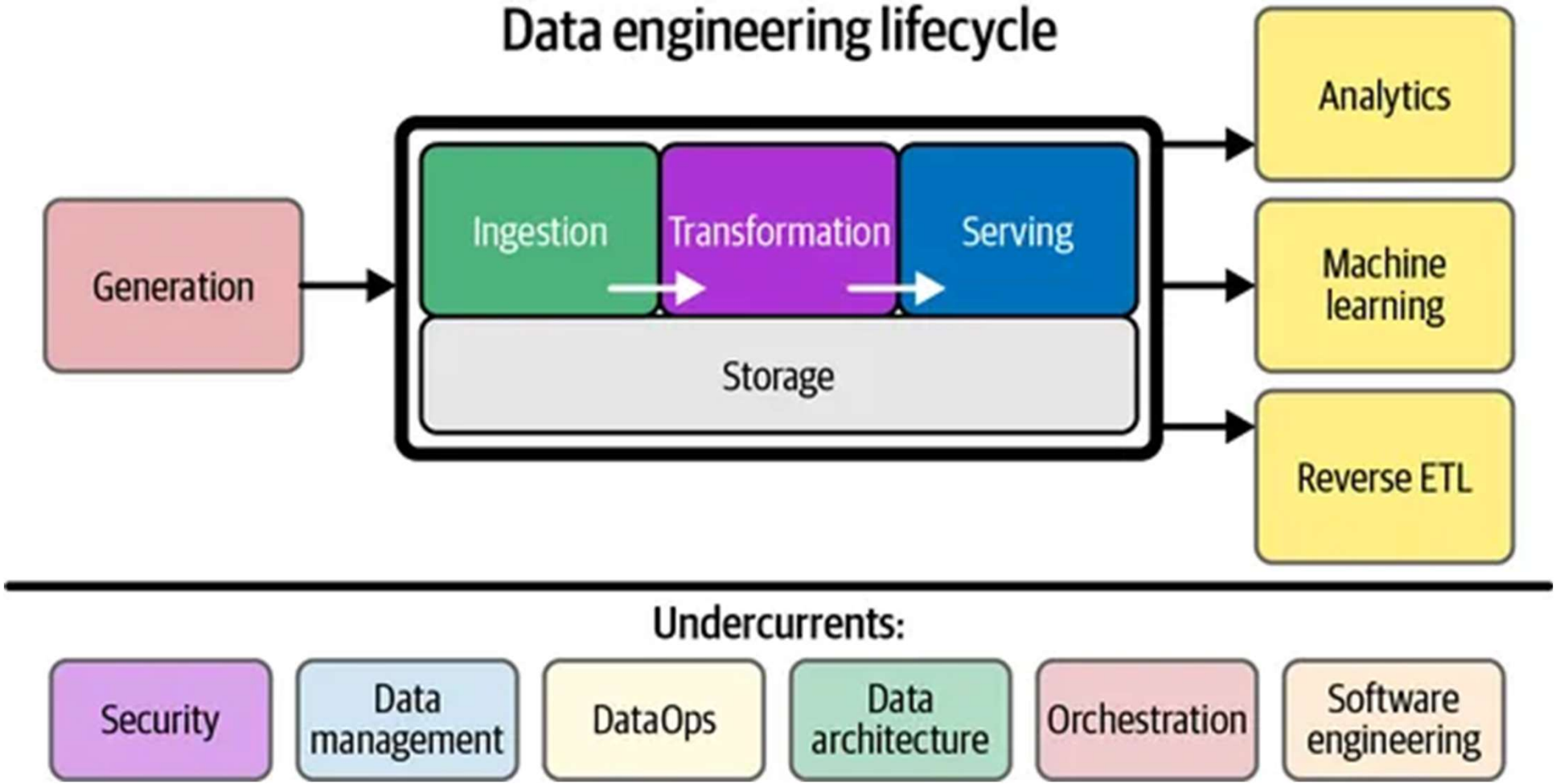
- Data scientists are the key to realizing the opportunities presented by big data. They find compelling patterns in data, and advise executives on the implications for products, processes, and decisions

Data Engineer

- Data engineers focus on managing and organizing data, building and maintaining databases and data pipelines

- Data engineers build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret.


Data Engineering Team
Data Modelling
Performance Tuning
Data Consultants
Data Custodians
Data Lake
Monitoring Infrastructure
Data Infrastructure
Data Pipelines
Data Warehouse Management
SQL
Database Experts


Data engineering lifecycle
Generation
Ingestion
Transformation
Serving
Storage
Analytics
Machine learning
Reverse ETL
Undercurrents:
Security
Data management
DataOps
Data architecture
Orchestration
Software engineering

Where is Data Engineering used?

- National Security
- Cyber Security
- Business Analytics
- Engineering
- Healthcare
- And more ….

Areas of Data Engineering

- Machine learning
- Data analysis
- ETL processes
- Build data systems and pipelines
- Data pipelines
- Data warehouse
- Security and compliance
- Build algorithms and prototypes
- Data cleansing
- Data security
- Quality Assurance

Areas of Data Engineering

- NoSQL
- Data accessibility
- Data modeling
- Data processing
- Database administration
- Architecture design
- Data analysis and synthesis
- Data integration
- Evaluate business needs and objectives
- Great numerical and analytical skills

Key Components of Data Engineering

1. **Data Collection**
   - **Data Sources**: Data engineers work with a variety of data sources, including databases, APIs, flat files, and real-time data streams.
   - **ETL (Extract, Transform, Load)**: This is the process of extracting data from different sources, transforming it into a suitable format, and loading it into a data warehouse or data lake.

2. **Data Storage**
   - **Databases**: Structured data is often stored in relational databases (e.g., MySQL, PostgreSQL) or NoSQL databases (e.g., MongoDB, Cassandra).
   - **Data Warehouses**: These are centralized repositories for storing large volumes of structured data.
   - **Data Lakes**: Used for storing vast amounts of raw data in its native format. Technologies include Hadoop HDFS, Amazon S3, and Azure Data Lake.

Key Components of Data Engineering

3. **Data Processing**
   - **Batch Processing**: Handling large volumes of data at rest. Tools include Apache Hadoop and Apache Spark.
   - **Stream Processing**: Real-time data processing. Technologies like Apache Kafka, Apache Flink, and Amazon Kinesis are commonly used.

4. **Data Integration**
   - **Middleware**: Tools and platforms that enable data integration across different systems and applications. Examples include Talend, Informatica, and MuleSoft.
   - **APIs**: Application Programming Interfaces facilitate the interaction between different software systems.

Key Components of Data Engineering

5 Data Quality and Governance

- **Data Quality**: Ensuring data accuracy, completeness, consistency, and reliability.
- **Data Governance**: Policies and procedures for managing data availability, usability, integrity, and security. Tools include Collibra and Alation.

6 Data Security

- **Encryption**: Protecting data at rest and in transit.
- **Access Control**: Managing permissions and user roles to ensure data is accessed by authorized personnel only.
- **Compliance**: Adhering to regulatory standards such as GDPR

Key Components of Data Engineering

7. **Monitoring and Maintenance**:
   - Regular monitoring and maintenance ensure data pipelines run smoothly. Engineers troubleshoot issues and adapt to changing requirements.
   - Data engineering solutions must scale with growing data volumes. Engineers optimize performance to handle increasing workloads

8. **Trends in Data Engineering**
   - **Cloud Computing**: Increasing adoption of cloud-based data infrastructure for scalability and flexibility.
   - **DataOps**: Applying DevOps principles to data management to improve agility and collaboration.
   - **Machine Learning Integration**: Embedding machine learning models within data pipelines for real-time analytics and decision-making.
   - **Real-Time Analytics**: Growing demand for real-time data processing and analytics to support dynamic business environments.

Key Components of Data Engineering

9. Tools and Technologies