Big Data Technologies

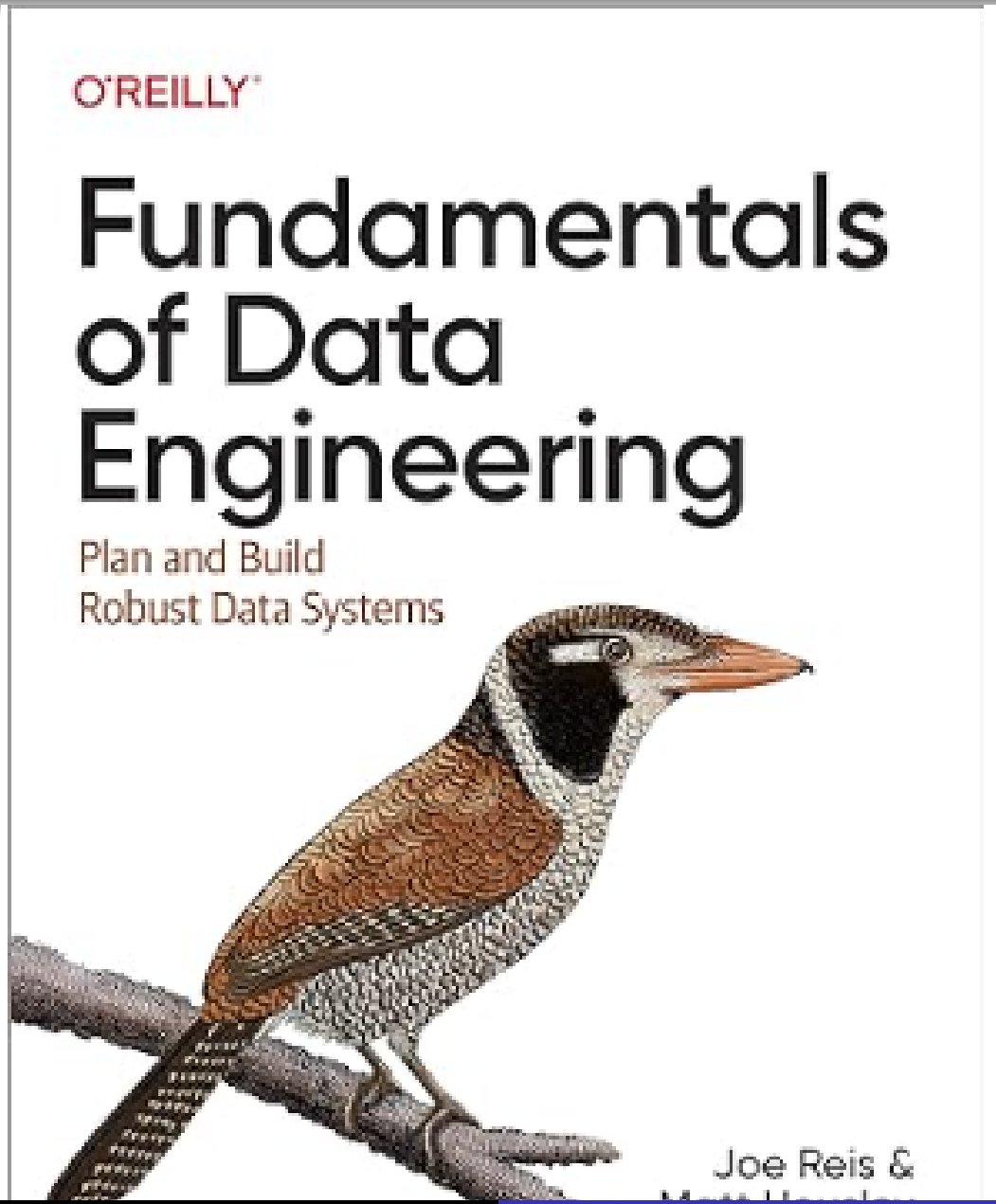
Chapter 01

Fundamentals of Data Engineering

Section 1

A Unified Analytics Engine

Text Book


O'REILLY
Fundamentals of Data Engineering
Plan and Build Robust Data Systems
Joe Reis &

Objectives

- Discuss and define the definitiondata engineering
- Describe the concept of Data Maturity and how to identify it
- Define and describe type A & B engineers
- Define and describe who data engineers work with

Outcomes

At the conclusion of this lecture and lab you have been introduced to day-to-day operation of a data engineer and the data lifecyle. You will be able to explain Data Maturity and how it impacts business and describe who a data engineer is and whom they work with.

This Semester's Outcome

- This class is a Big Data class – or more properly data class
- This class is a tooling class
- We will be covering the lower level process to setup tools and manage data
- We won't be covering explicit machine learning or mathematical formulas
    - You will apply what you learn here in your other courses

Small History of Big Data

- Which came first?
  - The relational model?
  - SQL Language?
  - Relational Database Software?

More History

- In the 1980's you see the creation of the RDBMS software
  - Michael Stonebreaker father of RDMBS
  - Ingress
- Oracle got their start around the same time (then known as Relational Software)

Data Engineering Described

- Exploring Data Engineering and what it is
- How was it born?
- Follow its evolution
- Define the skills a Data Engineer needs
- Show whom you will work with

Where Data Engineering comes from

- The term Big Data starts around 2010
  - The concept Data Engineering starts about 2020
- *"A data engineer manages the data engineering lifecycle, beginning with getting data from source systems and edning with servicn data for use cases, such as analytics or ML."*
  - *What is Data Engineering?*

Data Engineering LifeCycle

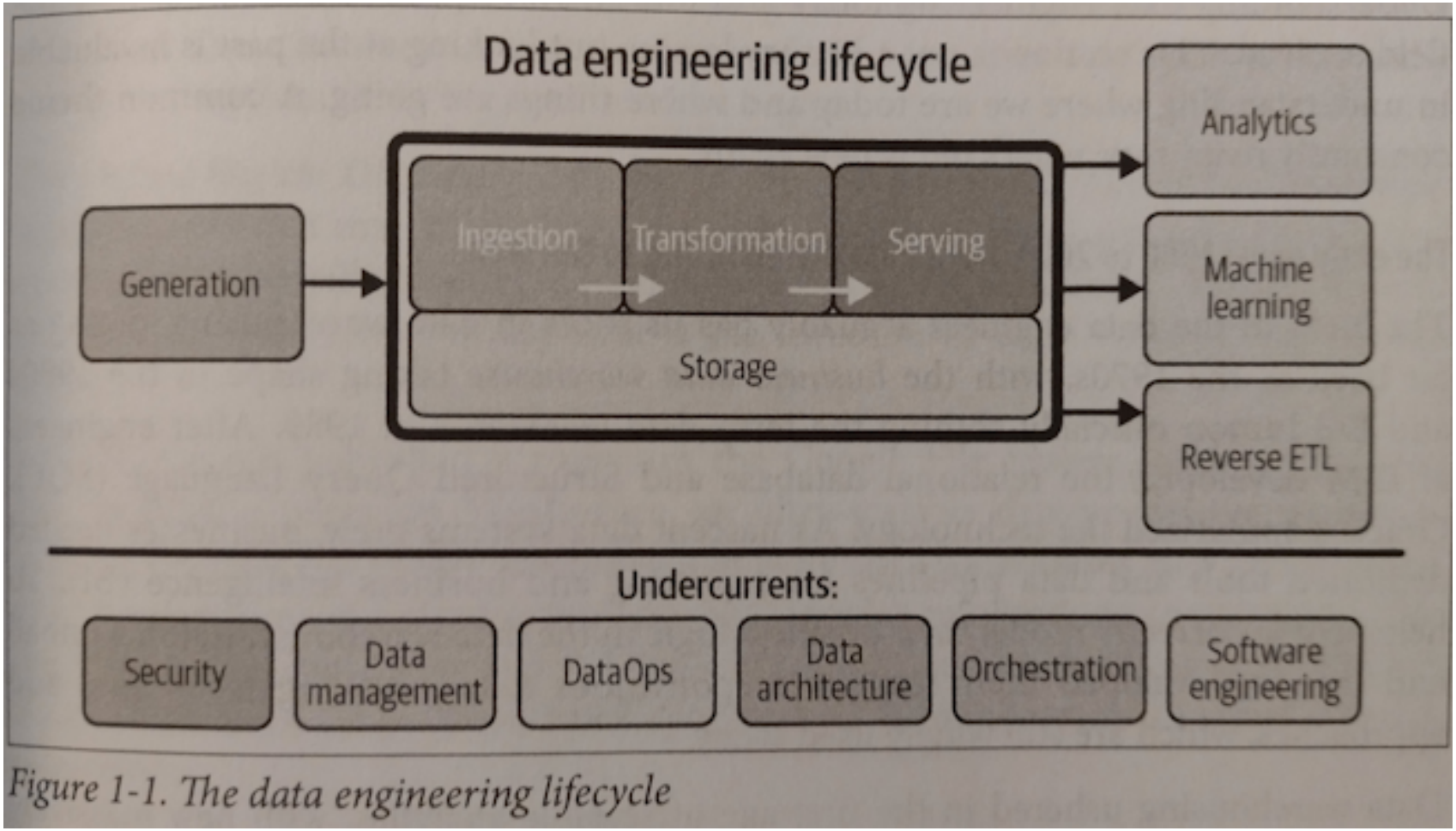


Figure 1-1. The data engineering lifecycle

Figure 2: *Data Engineering Lifecycle*

LifeCycle

- Shift our conversation away from tech stacks and on to concepts
- Stages of Data Lifecycle
  - Generation
  - Storage
  - Ingestion
  - Transformation
  - Serving
- Undercurrents of: Security, Data Management, DataOps, Data architecture, Orchestration, Software Engineering

History 1980 - 2000

- Data Warehousing to the web
- IBM creates and introduces SQL early 1980s
- Oracle popularizes/standardizes the use of SQL mid 1980s
- Bill Inmon
  - Coins term Data Warehouse in 1989
  - MPP - databases
  - Creates jobs such as BI Engineer, ETL Dev, Data Warehouse Engineer

Early 2000s

- Birth of Data Engineering
- Early e-com sites relied on Data Warehousing techniques
  - Not engineerd for speed/size of internet
- Cost of hardware went down as internet adoption/access went up
- Concept of Big Data was created
  - Concepts of MapReduce created

2000 to 2010

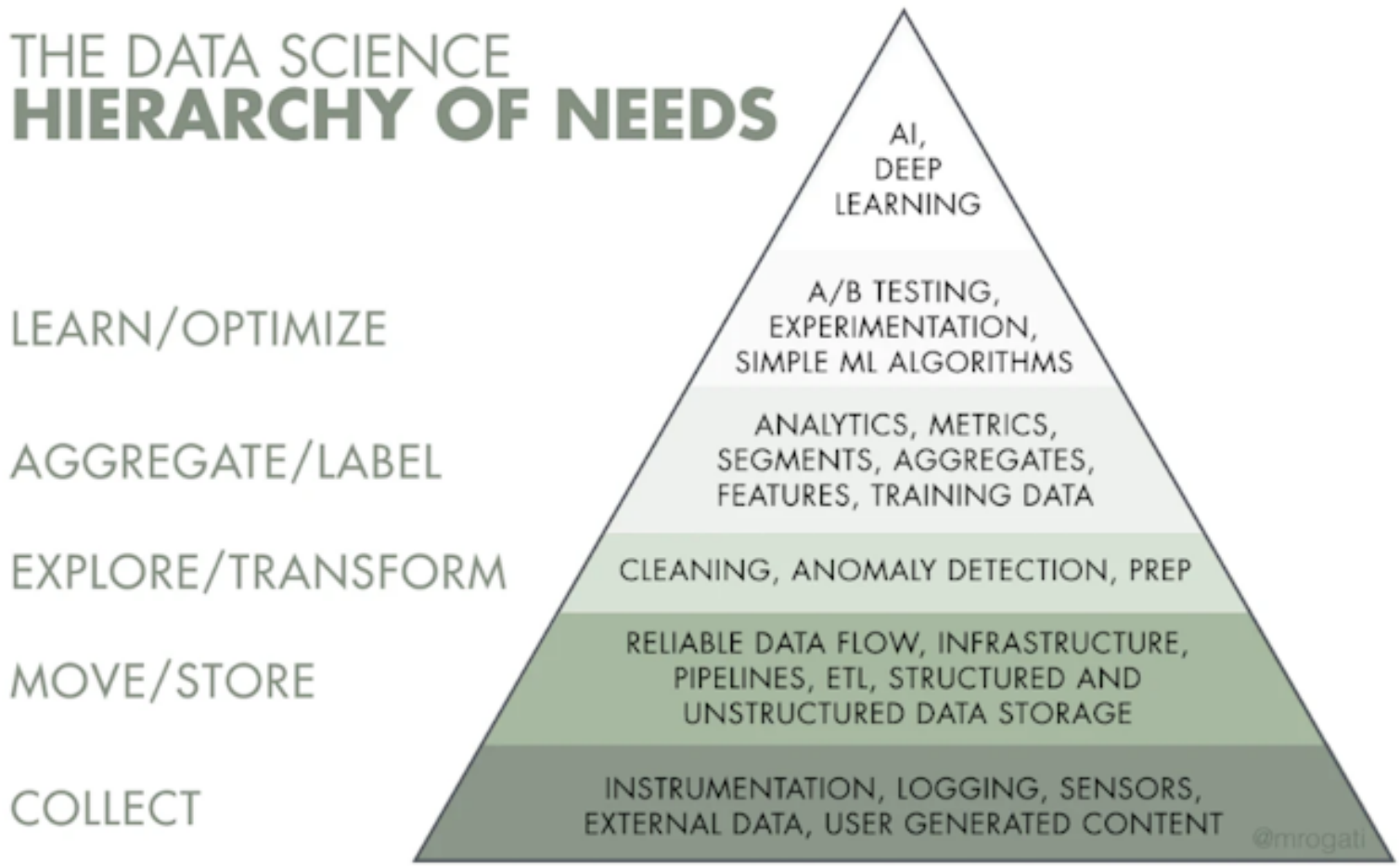
- Big Data Engineering
- Driven by Open Source
  - MySQL
  - Hadoop, Pig, Hive, HBase, Cassanandra, Presto
  - Many products from FaceBook
  - Amazon S3 (storage)
- Introduced complexity of management
- Realize *Big Data* is a relic of a term
  - Its all just Data and the Data LifeCycle

2020 and beyond

- *Data landscape image*
- Now focused on Data Engineering pipelines
  - Full LifeCycle
  - SQL is pretty much decided
- Worry about compliance and regulations for handling data
  - GDPR and other sovereign data protection laws

Data Science and Data Engineering Differences

- Data Sceince is downstream of Data Engineering
- *Hierarchy of Data Needs*



Data Engineering Skills and Activities

- Know how data is produces and consumed in the company
- We will cover in later chapters
  - cost
  - agility
  - scalability
  - simplicity
  - reuse
  - interoperability

Data Maturity

- Page 15 printed text, Three Phases
  - Start with data
  - Scale with data
  - Lead with data

Data Engineer Resposiblities

- Communicate with non-technical stake holders
- Understand the scope of business requirements
- Understand Agile, DevOps, and SRE terms and concepts
- Understand how to control costs and how data generates costs
- Learn Continually!
- Gain general skills
  - SQL, Python, Java Stack, Bash, and Linux

Types of Engineers

- Type A & B
  - Type A for application
    - Generally uses of the shelf software for solutions
  - Type B for builder
    - Tends to build solutions when needed internal to the company
- DE's sit in the middle of the developers/OPs and Analysts
  - SEs, Ops, Architects -> DEs -> Analysts, Scientists, ML/AI

Data Engineers talk to whom?

- CEO
- CIO
- CTO
- CDO
- CAO
- CAO-s (algorithms)
- Product and Project Managers
  - Page 28 and 29 printed

Conclusion

- We defined data engineering and the data lifecycle
- We described our data maturity model
- We described Type A & B engineers
- We described the DEs place in the corporate structure

Questions?

- Any questions?
  - Discord always open