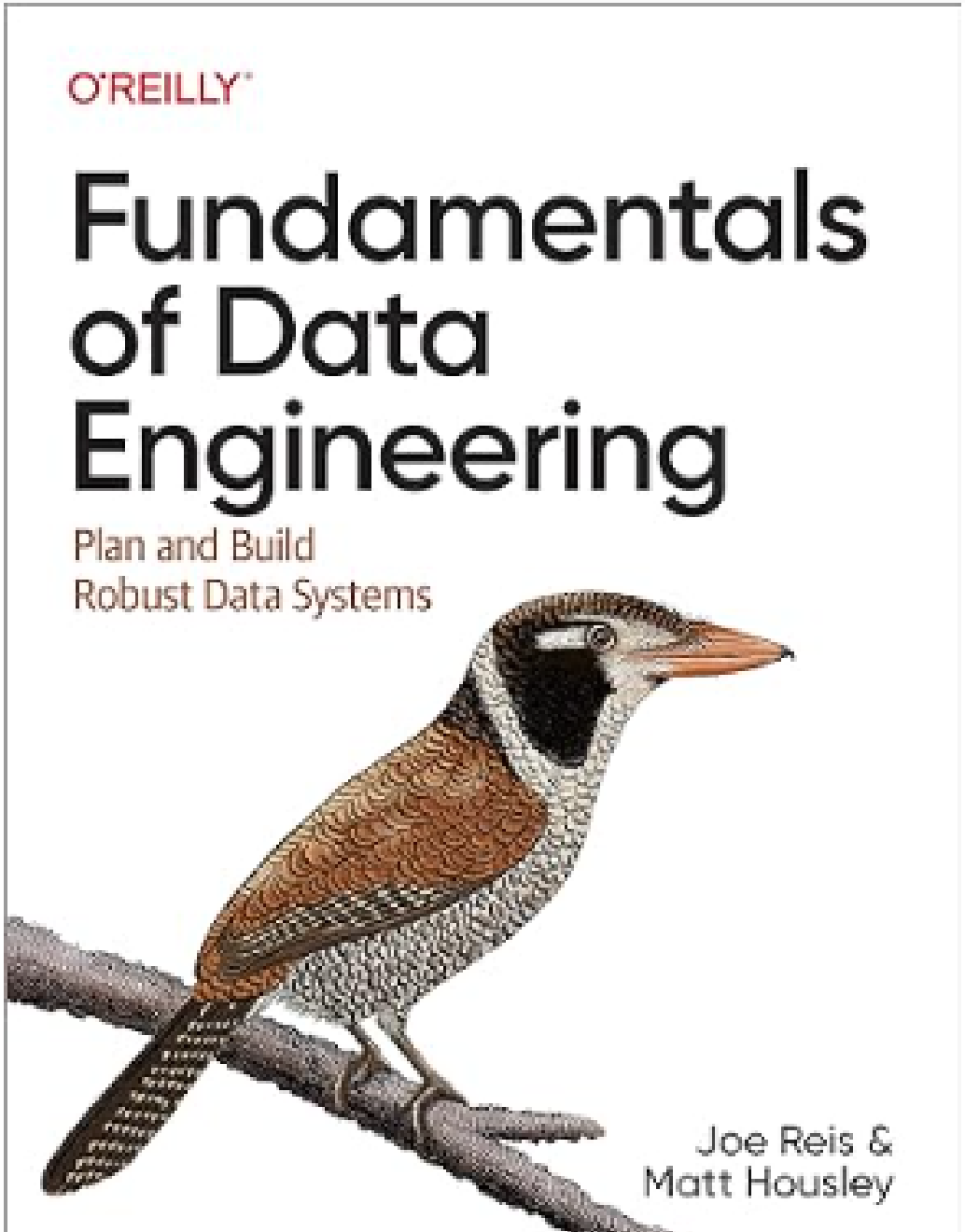
Big Data Technologies

Chapter 04

Fundamentals of Data Engineering

Choosing Technologies Across the Data Engineering Lifecycle

Text Book


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

Objectives

- Discuss tactical plan for making technology choices
- Explain the difference between architecture and tools
- Discuss the nature of software trade-offs and aspire for reversible decisions
- Explain TCO based decisions

Outcomes

At the conclusion of this lecture and lab you have covered the ranges of decisions that need to made when dealing with tooling choices. You will be able to express some of the trade-offs that need to be considered.

Review Chapter 1

- List the 5 stages of the Data Engineering lifecycle from Chapter 1?
- List the 4 technologies every data engineer should be familiar with?
- List the 5 business responsibilities of a data engineer
- List the 6 undercurrents to the Data Engineering lifecycle
- Which came first, Relational Database Model or SQL?

Review Chapter 2

- Define data life cycle management
- Discuss a Data Engineers relationship to business objectives

Review Chapter 3

- Defined data architecture
- Explain how data architecture sits fundamentally at the core of a business
- Explain 3 of the current data architectures

Choosing Technologies Across the Data Engineering Lifecycle

- Team size and capabilities
- Speed to market
- Interoperability
- Cost Optimization and business value
- Today vs the future
- Location (cloud, on prem, hybrid cloud, multicloud)
- Build verses buy
- Monolith vs modular
- Serverless verses servers
- Optimization, performance, and the benchmark wars
- The undercurrents of the data engineering life cycle

Team Size and Speed to Market

- What do you already have on your team?
  - Maximize those skills
  - Avoid copying other teams because they use the *latest* thing
- Speed to market wins
  - Your tech choices need to also enable you to get to market
  - If the software is not in the customers hand, it doesn't matter
  - Deliver value early and often

Interoperability

- Rarely will you use just one tech stack
  - Question of interop?
  - There is a reason the JSON and CSV files exist as standard data interchange formats
  - JDBC / ODBC
- Using REST APIs is universal, but not all APIs are the same
  - Not all compatible with each other

Cost Optimization and Business Values

- In a perfect world you would use the latest and the greatest
  - World isn't perfect
  - Organization expects an ROI from your projects
- Cost for us as Data Engineers measured through three methods
  - TCO
  - Opportunity Cost
  - FinOps

Total Cost of Ownership

- Includes direct and indirect costs
- Also known as overhead
  - CapEx
  - Opex
- Until the Cloud Era, OpEx wasn't really a choice

Total Opportunity Cost of Ownership

- The cost of choosing one way over all the other possibly ways
- What is the cost to move away from the system at a later point?

FinOps

- FinOps as Principle 9: Embrace FinOps from page 87
  - The goal is to fully operationalize financial accountability and business value
  - By deploying DevOps like thinking
    - Relating to deployment
    - Scaling
    - Monitoring
- *"FinOps is not about saving money. . . FinOps is about making money"*
  - Project like Kubecost
  - AWS Cost Explorer

Tech Today vs Tomorrow

- Think about what parts of technology will be around for the long haul
  - Immutable
  - Transitory
- Tech such as managed virtual machines will be around
  - EC2
  - VMWare
- Tech stacks like Object Storage will be around
  - AWS S3
  - Azure Blob Storage
- SQL
  - Generally not going anywhere
- Spark
  - SQL and Python

Transitory Tech

- JavaScript Frameworks
  - Backbone, Knockout, and Ember
  - React and Vue
- Data AI landscape
- Generally recommended to think in two-year frames
  - What will the tech landscape look two years from now?

Location - On Prem

- On prem vs Co-location vs Cloud
- On prem
  - CapEx
  - Advantages
    - Can tune your business as you have expertise
    - DropBox
    - 37 signals
    - Oxide.computer
    - Looking to build rack scale computing for on prem deployments
  - Data Gravity – where is your data?

Location - Cloud

- At first Public Cloud vendors pushed to get you there
  - At this point they have most of the workloads that can be moved/migrated
  - New avenues of business like OpenAI being run on Microsoft Azure
  - Newer features such as Lambda
    - Also known as serverless
    - Ability to spin up OS containers in a fraction of the time execute code and shut down the resource
    - Compute as a function
- Different economics
  - Need to apply FinOps

Location - Hybrid Cloud

- Using some on prem resources
- Using some cloud resources
  - Using Multicloud solutions
  - Prevents lock in to one cloud
  - Not all services are compatible
  - Need third party tools that are cross cloud
    - Hashicorp Terraform

Cloud Repatriation Arguments

- A Trillion Dollar Paradox
  - Large companies have a different scope of issues

Build vs Buy

- Common advice: buy where you can
  - Build requires an engineering culture
    - That can survive the loss of the key architects
  - Opensource
    - Also tricky
    - Companies can make use of all kinds of software and libraries available
    - Red Hat Linux turned opensource into a billion dollar business
    - You have the ability to contribute, inspect, and modify the code
  - Recently companies have pulled back from being Opensource
    - Fear that they are loosing control of project
    - Competitors can use their code in a different project (lost revenue stream)

Build vs Buy Commercial OSS

- Look for Commercial Opensource backing
  - Apache Spark is backed by the company Databricks
  - Apache Kafka is backed by Confluent
  - Can you self-manage?
  - Do you need to buy the enterprise product and support?
  - Some Cloud vendors may offer their own supported version
    - OpenSearch
    - DocumentDB

Conclusion

- Choosing technology isn't always easy
- Choosing technology is a balance
- Always be assessing trade-offs and reversibility

Additional Resources

- See page 155 and 156 of chapter three for additional readings

Homework

- Read FDOE chapter 5

Questions?

- Any questions?
  - Discord always open