Great Book
of
Data Engineering Concepts

About Me

My name is Khathiravan Raj Maadhaven. I have done Bachelor's in Computer Science and Engineering. I have been doing Softwares through-out my work life, with the Data Engineering as specialism since June 2010.

Thank you note Family support

<yet to be filled in>

Who is this book for? What to expect in this book?

I have learned a lot about Data Engineering concepts in a hard-way, I really wished there was one place where I could regularly go and check for new updates, every now and then.

The data is going through evolution phase and so the engineering around it too. The evolution of Big Data changed the face of parallel computing forever; with Cloud now into play, the legacy tera datasets problems have disappeared. With the legacy problems are disappearing, new patterns or new way of doing things are evolving.

Just like the TDD, BDD, DDD, Data Engineering is going through the same evolution phase with the MLOps and DataOps are becoming the next big thing.

This book is for someone who want to get an understanding of the Data Engineering Concepts without having to spend hours and hours on the internet to find them. Even when. I am collecting these concepts, I have picked up few new terms.

Now the most important part:

Not all of the content is not my own, I have collected / summarized from various articles or Blogs or even Books. The reference locations have been provided where possible.

Should you wish more details need to be added, please feel free to reach out to me.

Final note, the content or views expressed in this book are solely for the purpose of learning, not for profit reasons. These in no-way have any relation to businesses I am associated with (past, present and the future).

Contents

ABOUT ME ..... **- 2 -**
THANK YOU NOTE FAMILY SUPPORT ..... **- 3 -**
WHO IS THIS BOOK FOR? WHAT TO EXPECT IN THIS BOOK? ..... **- 4 -**
CONTENTS ..... **- 5 -**
EMERGENCE OF SPECIALISTS ..... **- 6 -**
INCREASE IN IMPORTANCE OF DATA & ENGINEERING ..... **- 7 -**
DATA ENGINEERING PRINCIPLES ..... **- 8 -**
DATA PIPELINES DESIGN PRINCIPLES ..... - 9 -
MIND THE KNOWLEDGE GAP ..... **- 11 -**
DATA BASED DECISION MAKING (DDDM) ..... **- 15 -**

Emergence of Specialists

As emergence of the specialists required for each Technical area, like UI/UX, Front End developer, Server side scripting, Middleware / API layer, service layers, back-end developer, ML/AI etc..,

Increase in importance of Data & Engineering

Data Engineering as specialist is coming into limelight now. As the evolution of the AI / ML requires quality data; in order to secure quality data, a Data Grid system is required.

Data Engineer – who? & role to business?

The main purpose of data engineer is to deliver an effective "Data Grid" system.
Data Engineering specialists not only understand the data solutions, they help in effectively design the flow of data and its architecture with a holistic view/control on all of the business data.

Data Engineering is the aspect of Data science that focuses on practical applications of data collection and analysis. For all the work that data scientists do to answer questions using large sets of information, there have to be mechanisms for collecting and validating that information.

- Provide guidance on the data flow design
- Controls the infrastructure requirements for the data
- Testing automation of data
- Setting up monitoring systems and maintenance of the data

Data Engineering teams are doing much more than just moving data from one place to another or writing transforms for the ETL pipeline. Data Engineering is more an umbrella term that covers data modelling, database administration, data warehouse design & implementation, ETL pipelines, data integration, database testing, CI/CD for data and other DataOps things.

Data Engineer – Role

Data Engineering Leader skills

- Should not only have tech skills
- Must have infrastructure mind
- Must have budget maintenance
- Must have team management
- Must be visionary to think of the next in line and the longer-term goal too

Data Engineering Principles

Let's not forget the starting a design is hard task, and even harder task is getting consensus on the architecture. Actually, there is quite a lot to take into consideration when designing data architectures, therefore having a proven set of principles would help you to prioritize and deliver.

https://blog.usebutton.com/3-design-principles-for-engineering-data

There are 3 foundational principles, listed by Jiaqi Liu. They are:

1. Design for Immutable Data
2. Create Data Lineage
3. Gradual and Optional Data Validation

Immutable Data

Immutable data is core to designing a system that is easy to test, that is idempotent and that is reproducible — without which the other two principles below are incredibly challenging to execute. Idempotent operations means that the same input will consistently produce the same output (no side effects).

As your code functionality becomes more and more complex — say instead of simple arithmetic, the code is running a machine learning algorithm — having immutable data that allows for reproducible testing becomes invaluable.

Creating Data Lineage

Data lineage on its own is the ability to trace and understand why data was mutated a certain way at a specific step in the pipeline.
Immutable data allows you to create data lineage.

On the architectural and data engineering side, there's a need to create observability for the data pipeline by tracing data lineage, which is identifying how data mutated at each iteration of the process. If at the end of a series of mutations, the outcome of the data is unexpected, you need to be able to iterate through each step of the pipeline in order to

identify at which step there was a either a bug in the code or an unexpected behavior of the data.

Gradual and Optional validation

It is the responsibility of data engineering to understand and document the numerical and data type assumptions (n > 0, field0: must be a string). Because it takes evidence to differentiate valid assumptions from invalid ones, it doesn't make sense to have strict validation at the front door.

This doesn't mean there's no validation. Instead, identify the non-negotiable security and privacy concerns around data and make sure those are handled with care. And then, allow for gradual and optional validation as you build up domain expertise about the data.

Without capturing the pieces of data that introduce uncertainty to the system, you can't readily determine what the correct behavior is. This means that validation has to be re-evaluated as data evolves and at a certain level, validation has to be optional. Ideally, the system can turn off bits and pieces of data validation to allow the processing of unexpected data.

With these principles followed, data systems can be built that are easy to test, idempotent, and traceable — all of which primarily leads to maintainability.

Combining these principles with Data pipelines design principles will form a formidable pair.

Data Pipelines design principles

Data Pipelines are at the centre of the responsibilities. To transform and transport data is one of the core responsibilities of the Data Engineer.

https://towardsdatascience.com/data-pipeline-design-principles-e7fbba070b4a

There are five qualities/principles that a data pipeline must have to contribute to the success of the overall data engineering effort.

Replayability

Irrespective of whether it's a real-time or a batch pipeline, a pipeline should be able to be replayed from any agreed-upon point-in-time to load the data again in case of bugs, unavailability of data at source or any number of issues. The feature of replayability rests on the principles of immutability, idempotency of data. This is what builds deterministicness into the data pipeline.

Auditability

For real-time pipelines, we can term this observability. The idea is to have a clear view of what is running (or what ran), what failed, how it failed so that it's easy to find action items to fix the pipeline. In a general sense, auditability is the quality of a data pipeline that enables the data engineering team to see the history of events in a sane, readable manner.

Scalability

It's a no brainier. Data is like entropy. It will always increase. To make sure that as the data gets bigger and bigger, the pipelines are well equipped to handle that, is essential. This would often lead data engineering teams to make choices about different types of scalable systems including fully-managed, serverless and so on.

Reliability

In addition to the data pipeline being reliable, reliability here also means that the data transformed and transported by the pipeline is also reliable — which means to say that enough thought and effort has gone into understanding engineering & business requirements, writing tests and reducing areas prone to manual error. A good metric could be the automation test coverage of the sources, targets and the data pipeline itself.

Security

In one of his testimonies to the Congress, when asked whether the Europeans are right on the data privacy issues, Mark Zuckerberg said they usually get it right the first time. Data privacy is important. GDPR has set the standard for the world to follow. Most countries in the world adhere to some level of data security. To have different levels of security for countries, states, industries, businesses and peers poses a great challenge for the engineering folks.

To make sure that the data pipeline adheres to the security & compliance requirements is of utmost importance and in many cases it is legally binding. Security breaches and data leaks have brought companies down. It's worth investing in the technologies that matter.

Mind the Knowledge Gap

Business and Tech people generally seem to speak different and somewhat specialized languages. Even when we use the same terms, it is often with a different meaning.
More than ever before, it is becoming absolutely important that Tech and business working relationships become co-operative and collaborative where communication is going to play vital role. Even so, the cross-skilling of Tech learning to become more business skills and business becoming Tech skilled will go a long way into bridging the gap between these two. Bridging this gap is the first step towards the data-based decision making.

In simple terms, Tech people must become business people and the Business people must become Technology people.

First and foremost, getting an understanding will provide us great in-depth details on where our focus should be. The best way to achieve an understanding is by doing evaluation / assessment. Best way to do evaluation is by preparing a questionnaire or a survey. There are many ways one can prepare this evaluation / assessment; I found the ten-question survey designed by Eckerson Group as best and useful starting point.

https://www.eckerson.com/articles/bringing-business-and-it-together-an-imperative-for-data-driven-business

Tech teams must have an understanding on the following, in order to develop, deploy and sustain high-impact information and Analytics systems. At the same time, they develop an appreciation for the challenges and complexities in the business management. As illustrated in the following, the answers for each of questions must be communicated against every domain.

Domains	**vs**	**Questions for each domain**
Business performance management		What is Business process management
Customer relationship management		Who performs Business process management?
Supply chain management		Why perform business process management?
Financial management		What are the key metrics of business process management?
Human resources management		What are the major activities of business process management?
Operations management		What are the issues and challenges of business process management?
Any other management domains		How does business process management support the mission and goals of the business?

Tech is equally responsible for preparing a Technical Stack Catalogue which details all the technology the information systems depend on, and the dependencies. Business people understanding on these would make huge impact in the capability to gather information, perform business analysis and so the informed business decisions.

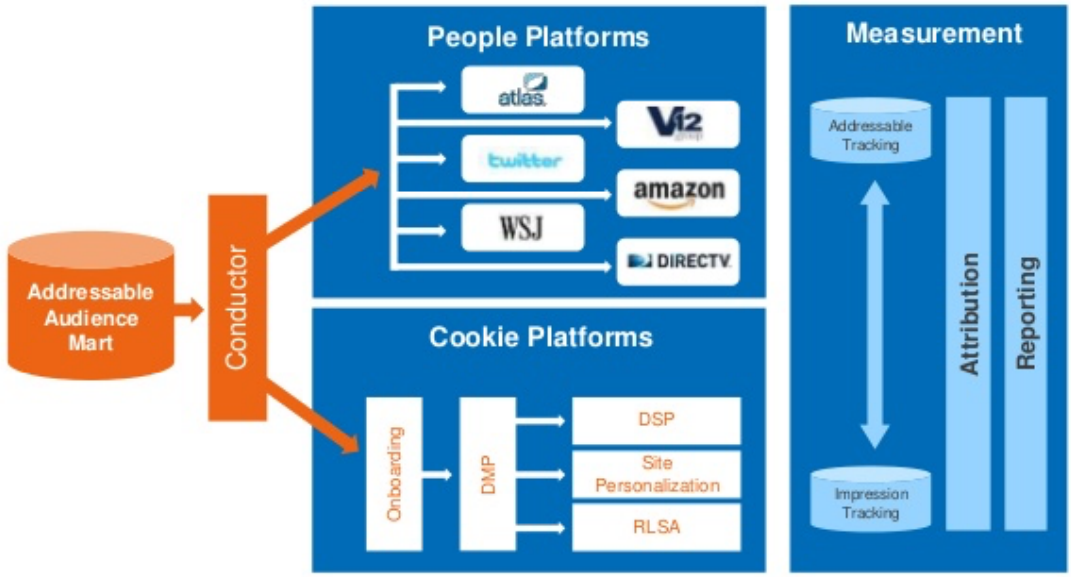
The following table provides a list of tech stack which can be extended and the sample tech stack illustration diagram.

Technology Stack Table

Data Science	Artificial Intelligence, Machine Learning, Automation
Business Analytics	Descriptive, Diagnostic, and Predictive Analytics
Performance Management	Dashboards, Scorecards, Business Metrics
Data Resources and Services	Data Catalogs, Data Lakes, Data Warehouses
Data Management Systems	DBMSs, Integration, Preparation, Quality, Lineage
Data Sources	Operational Systems, Big Data Sources, Partner Data Sources
Data Storage	DAS, SAN, Cloud, etc.
IT Administration Systems	Scheduling, Monitoring, Management, Tuning
Middleware	Communications, Messaging, Replication
Networks	LAN, WAN, Intranet, Internet
Operating Systems	Windows, Unix, Linux, etc.
Hardware / Compute Platforms	Servers, Desktop, Cloud, etc.

<insert a picture of Tech Stack example>

A mid funnel data and technology flow supports the the connectivity of the 3Cs



MERKLE

For each layer of technology, the following questions needs to be answered to have greater in-depth understanding:

- What role or purpose does the technology serve?
- What can I do when the technology works?
- What can't I do when the technology fails?
- What other technologies does it depend upon?
- What are the difficulties and risks of technology?

Finally, to close the gap, there are certain actions taken by Technology and Business to work together.

Action – owner who?	Tech	Business
Maintain good business/IT relationships at the line management level.	Yes	No
Create a federated IT culture	Yes	No
Operate transparent IT programs	Yes	No
Define and communicate technology strategy	Yes	No

Align Technology with Business processes	No	Yes
Embrace the right technology for the right reasons	No	Yes
Define and communicate the business strategy and vision	No	Yes
Align strategically in two directions	Yes	Yes
Align tactically in two directions	Yes	Yes
Align periodically in two directors	Yes	Yes

There are a lot of frameworks available for Organizational alignment, these techniques are not going to be covered here.

Data based decision making (DDDM)

With the Tech and Business gaps are bridged or figured out a way to move forward, the data-based decision making should become the part of it. This is called “Data Driven Culture”, as coming together of different functions to form a collaborative team to achieve the best using data.

Over the last few years, the importance of data have grown massively. As the IoT systems continue to grow the data, the importance of fact-based decisions proven to be more accurate than the experience based / theory based.

Data-driven decision-making (sometimes abbreviated as DDDM) is the process of using data to inform your decision-making process and validate a course of action before committing to it.

Why data driven decision making important?

Importance of data in decision lies in consistency and continual growth. It enables companies to create new business opportunities, generate more revenue, predict future trends, optimize current operational trends, and produce actionable insights.

https://online.hbs.edu/blog/post/data-driven-decision-making

How exactly data can be incorporated into the decision-making process will depend on a number of factors, such as your business goals and the types and quality of data you have access to.

The collection and analysis of data has long played an important role in enterprise-level corporations and organizations. But as humanity generates more than 2.5 quintillion bytes of data each day, it's never been easier for businesses of all sizes to collect, analyze, and interpret data into real, actionable insights. Though data-driven decision-making has existed in business in one form or another for centuries, it’s a truly modern phenomenon.

Pros of DDDM

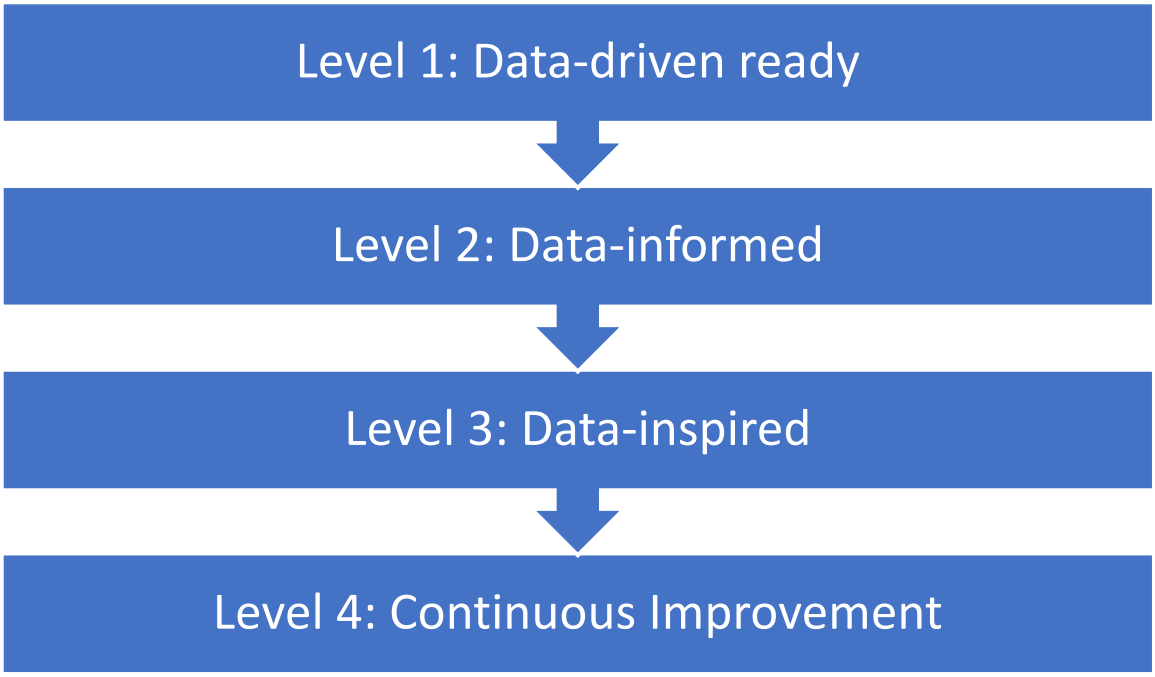
1. You’ll make more confident decisions
2. You’ll become more proactive
3. You can realize cost savings

To achieve, data driven you’ll have to:

1. Look for patterns everywhere
2. Tie every decision back to data
3. Continual learning

Data driven maturity levels

In order to achieve DDDM successfully, each organisation will have to go through 3 levels, with the last stage is catching up with the latest trends.

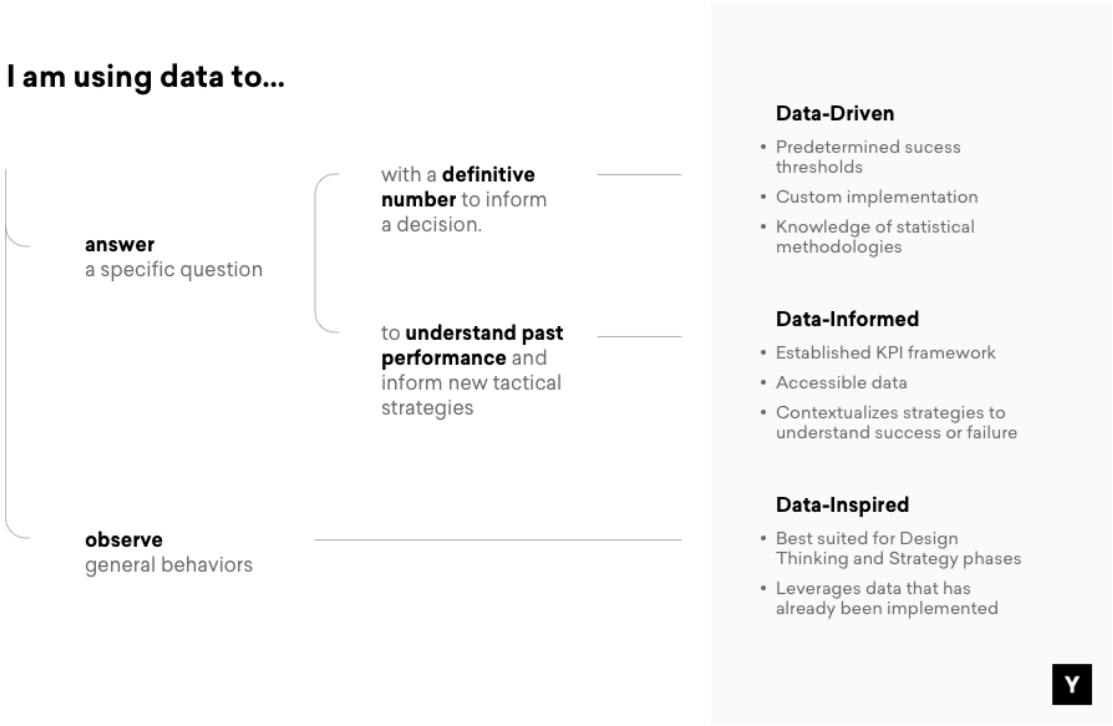


Data-driven ready – a level achieved only when the exact data for the business decisions are collected.

Data-informed – a level achieved only when the data is put into good use for understanding the current performance and behaviour.

Data-inspired – a level achieved only when the data is used to do trendspotting and predicting future customer expectations / peak usage etc..,

https://blog.amplitude.com/data-driven-data-informed-data-inspired



Identifying the level of data-driven decision level in any business, will help in getting the maximum out of the available data. The data-driven decisions systems are not CAPEX / one time investment, there is OPEX involved in the form of “Continuous Improvement”.

According to Gartner survey on the Supply-Chain, approximately 70% of the data gets wasted.

Data Engineers must make sure the data collected are actually going to be used, will be used. The systems design should accommodate the turning on of data points collection with minimal effort or work involved, the same applies for turning off of certain data points.