Real-World Data Engineering

Vicente Ruben Del Pino Ruiz
Agenda

ABOUT ME
DATA ENGINEERING
KEY COMPONENTS
CHALLENGES
KEY TAKEAWAYS
About me

Vicente Ruben Del Pino
(https://www.linkedin.com/in/vrdelpino/)

More than a decade working with data
- Cloud/Premise
- Big Data / NoSQL / RDBMS
- Several patents in data processes
- Leading Data teams

Disclaimer:
- Personal point of view
  - What has worked for me
  - What didn’t work for me
What is Data Engineering?
Data Engineering - Goals

Provide Data

- Timely Manner: Time to data
- Reliable: Right Data, Traceable, Automated, Production Env
- Usability: User Friendly, Right Info
- Secure: Right People, Right Data
- Best Practices: Right Methodologies

Right People
Right Data
Secure
Data Engineering - Skills

Which skills define a Data Engineer.

- **Data Modelling**: Understand and apply data modelling techniques
- **Data Processing**: Understand and apply data processing techniques
- **Software Engineering**: Coding skills, Automation, CI/CD
- **Expert in everything**: Beware of recruiters or teams where they ask for everything
All is about data.
Key Components

Key components to be included in any data product generated by Data Engineering area.

<table>
<thead>
<tr>
<th>Data Quality</th>
<th>Data Lineage</th>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right data set</td>
<td>Source of data</td>
<td>Information contained</td>
</tr>
<tr>
<td>How data is generated</td>
<td>System Owner</td>
<td>Meaning</td>
</tr>
<tr>
<td>Quality of data</td>
<td>Frequency</td>
<td>Right data set</td>
</tr>
<tr>
<td></td>
<td>Dependencies</td>
<td>Frequency</td>
</tr>
</tbody>
</table>
Key Components – Data Quality

**Proactive**
Detect issues and address them before users do:
- Alert System
- Smart Data Processing

**Informative**
DQ information available through:
- Data Mart
- Reporting Suite

**Automatic**
Any new data source included automatically
No manual intervention
Key Components – Data Lineage

**Full Coverage**
Cover starting from source system generating the data to end point of the pipeline

**Public**
Available to everybody
Graph visualization
Search engine

**Automatic**
Any new data source included automatically
No manual intervention
Key Components – Metadata

**Full Coverage**
Cover any data source
Cover any table and database

**Information**
Naming convention
Descriptions
Meanings

**Automatic**
Any new data source included automatically
No manual intervention
Challenges

1. Funding secured for an initiative or project. Building Data Engineering team.
2. First initiatives agreed. Scope and Goals defined.
3. Success from past projects ignite interest from business. More initiatives come on board. Prioritization among different projects.
Challenges – Where is my data?

Funding and budget approved.
Stakeholders / Business won't wait 6 months.

Borrow Team Members
Use team members from other projects to spin up the new team faster.

Infrastructure setup
Allocate some team members to setup the infrastructure.

Right People!
Focus on hire the right people. Don’t hire out of necessity.
Challenges – Architecture/Technology

Decide the right architecture for the area.
Right technology for the job.

Use Case

Decide your architecture design based on data consumption and generation.

Organization Architecture

Keep a close eye to recommended architectures in your organization.

Team skills set

Use the design that maximizes the usage of your team skills.
Challenges – Right members for the team

Competitive Market
Recruiters/Offers asking for all skills in one person.

Build on right skill set
Write down skills your team needs. Build a group of people (not a person) that covers all of them.

Hire the right mix
Hire senior and junior. At least 1 senior per 2 juniors (beware not all seniors want to mentor).

More time
Learning and upskill process takes time.
Challenges – Requirements

Lack of documentation in sources.
No information about data access.

- **Data Governance**
  Data Governance to point to the right sources and the SME for data sets.

- **Data Quality**
  Data Quality to provide information about quality in the data source.

- **Metadata**
  Metadata to know what is in there.
Challenges – Prioritization

Beware of success.
There is never enough data.

Data Governance
Decide which one is the right data set to process first.

Team size
Plan the right size of the team for the number of projects.

Technical Debt
Avoid Technical Debt. Each time you cut corners, you are stealing your future team capacity. Allocate resources to pay the debt.
Challenges – Automation

Not enough capacity or time to build generic processes.
Capacity taken by manual/repetitive processes

- Build Generic Configuration driven
  Build no custom pipeline or process. All processes generic and driven by configuration.

- Automate everything
  Any task repeated more than once should be automated.

- Automatic Releases
  Build automatic release pipelines to any environment.
Challenges – Technical Debt

High priorities and critical tasks forcing the team to generate technical debt.
No time available to pay back the technical debt generated.

- Don’t cut corners
  Avoid short-term developments.
  Think always long-term.

- Build generic first
  Build generic processes driven by configuration.

- Automate, automate, automate
  Any repetitive task should be automated. Always.
Challenges – Support

Once processes are in Production, they need support. Everything will eventually fail, be ready for it.

Allocate capacity for support

Ensure you have room for support anything failing in your processes.

Build an alert system

Build a system to alert when something doesn’t go as expected.

Build easy to re-run processes

Implement best practices for easy to re-run processes.
Takeaways

**DQ/DL/DG**
These are the three key pillars for any data area to be resilient and reliable

**Automation**
Automate, Automate, Automate
Anything not automated will take capacity and resources from you in the future.

**Talk with me**
If you want to talk about Data connect with me in LinkedIn
Thank You!

Q&A